S H E H U I K E X U E Y A N J I U F A N G F A J I N G D I A N Y I C O N G

FUZA DIAOCHA
SHEJI YU FENXI
DE SHIYONG FANGFA


社会科学研究方法经典译丛
SHEHUI KEXUE YANJIU FANGFA JINGDIAN YICONG

复杂调查设计与分析的实用方法

■里斯托·雷同能　厄尔基·帕金能 著
■王天夫 译


重庆大学出版社
http://www.cqup.com.cn

S H E H U I K E X U E Y A N J I U F A N G F A J I N G D I A N Y I C O N G

FUZA DIAOCHA
SHEJI YU FENXI
DE SHIYONG FANGFA

研究人员在探讨科学问题的时候，经常发现需要分析复杂调查数据。要恰当地分析数据，了解复杂调查设计的不同方面是相当重要的。《复杂调查设计与分析的实用方法》特别涵盖设计和分析复杂调查的中高级统计技术。

这一新的版本完整地介绍描述性调查的抽样和估算，详尽地讲解复杂调查的分析，包含了许多新材料及实际生活中的例子。

本书涵盖了商务、教育、卫生、社会经济的调查与官方统计方法，适用于计划、实施和分析复杂调查和民意调查的研究人员和实际工作者。同时，本书也适合作为中高级抽样调查课程的教材。

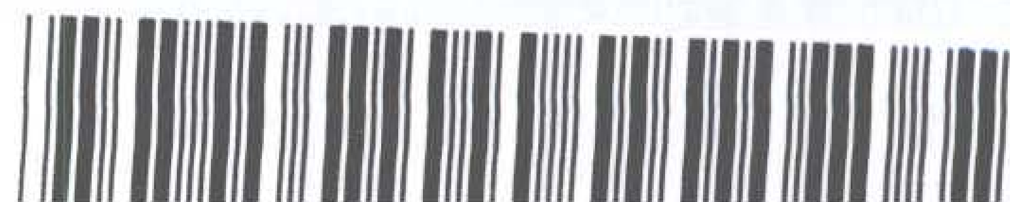
万卷方法博客圈:
http://q.blog.sina.com.cn/fafang

ISBN 978-7-5624-4290-5

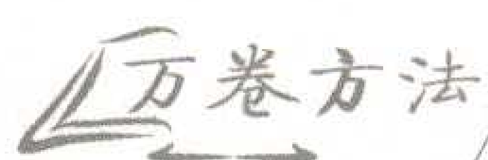


定价:45.00元

S H E H U I K E X U E Y A N J I U F A N G F A J I N



FUZA DIAOCHA
SHEJI YU FENXI
DE SHIYONG FANGFA



社会科学研究方法经典译丛
SHEHUI KEXUE YANJIU FANGFA JINGDIAN YICONG

■主编 沈崇麟 夏传玲

复杂调查设计与分析的实用方法

■里斯托·雷同能 厄尔基·帕金能 著
■王天夫 译

重庆大学出版社

图书在版编目(CIP)数据

复杂调查设计与分析的实用方法/(芬)雷同能(Lehtonen, R.),(芬)帕金能(Pahkinen, E.)著;王天夫译.—重庆:重庆大学出版社,2008.1
(万卷方法.社会科学研究方法经典译丛)
书名原文:Practical Methods for Design and Analysis of Complex Surveys
ISBN 978-7-5624-4290-5

Ⅰ.复… Ⅱ.①雷…②帕…③王… Ⅲ.调查研究—方法 Ⅳ.C31

中国版本图书馆 CIP 数据核字(2007)第 165225 号

复杂调查设计与分析的实用方法
里斯托·雷同能 厄尔基·帕金能 著
王天夫 译
责任编辑:雷少波 罗 杉 版式设计:雷少波
责任校对:刘雯娜 责任印制:张 策
*
重庆大学出版社出版发行
出版人:张鸽盛
社址:重庆市沙坪坝正街174号重庆大学(A区)内
邮编:400030
电话:(023) 65102378 65105781
传真:(023) 65103686 65105565
网址:http://www.cqup.com.cn
邮箱:fxk@cqup.com.cn(市场营销部)
全国新华书店经销
自贡新华印刷厂印刷
*
开本:787×1092 1/16 印张:18 字数:342千
2008年1月第1版 2008年1月第1次印刷
印数:1—3 000
ISBN 978-7-5624-4290-5 定价:45.00元

万卷方法

总策划：崔 祝 雷少波

调查研究方法（第3版）
弗洛德·J·福勒 著 孙振东 等译

量表编制：理论与应用（第2版）
罗伯特·K·德维利斯 著 魏勇刚 等译 李红 校

案例研究：设计与方法（第3版）
罗伯特·K·殷 著 周海涛 等译

案例研究方法的应用（第2版）
罗伯特·K·殷 著 周海涛 等译

解释性交往行动主义：个人经历的倾听、叙事与理解（第2版）
诺曼·K·邓金 著 周勇 译

电话调查方法：抽样、筛选与监控（第 2 版）
保罗·J·拉弗拉卡斯 著 沈崇麟 译

科学决策方法：从社会科学研究到政策分析
罗格·沃恩 著 沈崇麟 译

研究设计与社会测量导引（第6版）
迪尔伯特·C·米勒 著 风笑天 等译

公共管理定量分析：方法与技术
袁政 著

论教育科学：基于文化哲学的批判与建构
申仁洪 著

复杂性科学的方法论研究
黄欣荣 著

社会科学研究：方法评论
陈向明 等主编

公共政策内容分析方法：理论与应用
李钢 等编著

质化方法在教育研究中的应用：个案研究的扩展
莎兰．B．麦瑞尔姆 著 于泽元 译

复杂调查设计与分析的实用方法
里斯托·雷同能 厄尔基·帕金能 著 王天夫 译

质性资料的分析
Matthew B. Miles A. Michael Huberman 著 张芬芬 译

社会研究方法
仇立平 著

美国心理学会写作手册（第5版）
美国心理学会 编 陈玉玲 王明杰 译

万卷方法

研究设计与写作指导：定性、定量与混合研究的路径(第2版)
约翰·克雷斯威尔 著　崔延强 等译　孙振东 校

社会网络分析法(第2版)
约翰·斯科特 著　刘军 译　沈崇麟 校

定性研究(第1卷)：方法论基础(第2版)
诺曼·K·邓津 主编　风笑天 等译

定性研究(第2卷)：策略与艺术(第2版)
诺曼·K·邓津 主编　风笑天 等译

定性研究(第3卷)：经验资料收集与分析的方法(第2版)
诺曼·K·邓津 主编　风笑天 等译

定性研究(第4卷)：解释、评估与描述的艺术及定性研究的未来(第2版)
诺曼·K·邓津 主编　风笑天 等译

组织诊断：方法、模型和过程(第3版)
迈克尔·I·哈里森　张小山 译

民族志：步步深入(第2版)
大卫·费特曼 著　龚建华 译

分组比较的统计分析
廖福挺 著　高勇 译　沈崇麟 校

抽样调查设计导论(第2版)
罗纳德·扎加　约翰尼·布莱尔 著　沈崇麟 译

焦点团体：应用研究实践指南(第3版)
理查德·A·克鲁杰　林小英 译

质的研究的设计：一种互动的取向(第2版)
约瑟夫·A·马克斯威尔 朱光明 译　陈向明 校

多层次模型分析导论(第2版)
Ita kreft　Jan De Leeuw 著　邱皓政 译　郭志刚 校

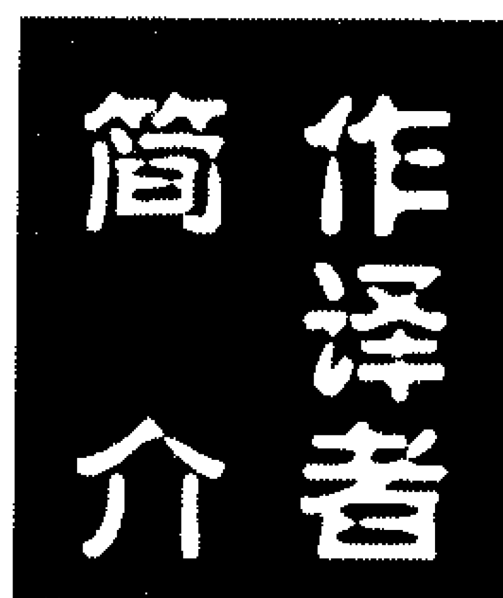
评估：方法与技术(第7版)
彼得 罗希 等著　邱泽奇 等译

实用数据再分析法
马克·利普西 著　刘军 译

叙事研究：阅读、分析和诠释
利布里奇 等著　王红艳 译

爱上统计学（第2版）
内尔·J·萨尔金德 著　史玲玲 译

哲学史方法论讲演录
邓晓芒 著



里斯托·雷同能(Risto Lehtonen)　赫尔辛基大学数学与统计学系教授,主要研究方向为抽样调查与统计分析方法。

厄尔基·帕金能(Erkki Pahkinen)　芬兰基瓦斯开拉大学数学与统计学系退休教授。

王天夫　2004 年于芝加哥大学获社会学博士学位,现任职于清华大学社会学系。研究兴趣包括社会不平等、城市化进程、家庭与婚姻社会学、社会研究方法与统计分析技术等。

总序 社会研究方法的现状及其发展趋势

近年来，社会调查技术和社会研究方法都有很大的发展。在调查技术方面，自20世纪70年代以来，社会变迁多次横断面的跟踪调查研究，几乎成为所有国家和地区了解社会结构转变和社会发展状况的基础性调查。这种调查不仅对社会学的研究有很大促进，对整个社会科学的研究都产生了重大影响，而且这些调查结果有的已成为政府有关部门决策的重要依据。国际上比较著名的此类调查有：美国芝加哥大学全国民意调查中心（National Opinion Research Center，简称NORC）的“社会综合调查（General Social Survey，简称GSS）”，英国埃塞克斯大学调查中心进行的“全国家庭生活和社会变迁调查”，法国经济和社会调查所进行的“全国经济社会调查”，日本社会学会组织进行“全国社会分层与社会流动调查（简称SSM）”。中国台湾“中央”研究院社会学研究所，也每两年进行一次“台湾社会变迁基本调查”。美国的“社会基础调查”，现在已成为年度性的调查项目，它是美国国家基金会目前资助的最大的社会科学研究项目。以上这些调查，除美国的调查外，一般均因经费原因采用纵向的间隔性重复调查法，即每隔一段时间，进行一次全国规模的抽样调查。每次调查除保留社会研究所需的基本项目外，每次都有不同的主题。在间隔若干时间后，再重复同一主题的调查，这样的研究设计，使社会变迁研究在可以涉及更为广泛的研究领域的同时，具有更好的积累性和可比性。多年来，这些基础性调查获得的资料，滋养着大批的社会科学研究者，有时一项调查就有很多名博士生用来写博士论文，以此所取得的研究成就，其可靠性受到社会科学界的广泛认同。例如1997年出版的，以台湾社会变迁基本调查数据为基础的研究报告集《90年代的台湾社会，社会变迁基本调查研究系列二》收论文16篇，内容涉及社会生活的各个方面，在台湾引起了极大的反响。

国内社会科学界在这方面也有了长足的发展。笔者所在的中国社会科学院社会学研究所的社会调查和方法研究室，组织或参与了多项与社会变迁有关的大规模抽样调查，取得了一定的研究成果，并积累了大量有关社会变迁的宝贵数据资料，其中主要有：

1. 城乡家庭变迁系列调查：该课题是由中国社会科学院社会学研究所牵头，联合北京大学和地方社科院的研究人员展开的一项类似多次横断面的城乡家庭变迁调查。这一调查始于1981年的“中国五城市婚姻家庭调查”，而后有1988年的“中国农村家庭调查”、1991年的“中国七城市家庭调查”、1998年的“中国城乡家庭变迁调查”。

2. 有关中国城乡社会变迁的系列调查：调查始于1991年的第二批国情调查，

然后有1992年的“中国城乡居民生活调查”、1993年的“第三批国情调查”、1995年的“第四批国情调查”和1997年的“中国沿海发达地区社会变迁调查”。上述调查虽然还不是严格意义上的多次横断面的纵贯研究，但研究者已在研究设计中尽量考虑到纵贯研究的基本原则，如调查队伍的稳定、指标的可比性和样本空间的延续性等。

3. 中国城乡社会变迁调查：这一调查开始于2000年，为中国社会科学院重大课题。目前已经完成第一期第一次调查和第二次调查，今后将把这一调查发展为连续的、定期进行的社会变迁调查。

在纵向调查技术取得长足进步的同时，上世纪末至今，电话调查技术也有很大发展。电话调查涉及的范围几乎与个别（面对面）访谈同样全面。电话调查中使用的一系列方法，是在20世纪70年代后期和面对面调查一起发展起来的。在20世纪80年代中，电话调查开始变得很普遍，且成为许多场合中各种调查方法的首选。正如某些学者所言，一种在公共和私营部门被人们用来帮助提高决策效率的收集信息的有效方法为人们所普遍认同时，这一现象本身就具有方法论上的意义。不仅如此，电话调查还有很大的实践意义，因为它为研究者提供了更多的控制调查质量的机会。这一机会包括抽样、被调查人的选择、问卷题项的提问、计算机辅助电话访谈（CATI）和数据录入。正因为如此，今天在各种社会调查中，如果没有发现其他重要的足以放弃使用电话调查的原因，电话调查由于其独特的对调查质量进行全面监控的优点，常常成为各种调查方式的首选。由笔者翻译，重庆大学出版社出版的《电话调查方法：抽样、选择和督导》一书，也于2005年面世。

无论是纵向调查抑或电话调查，实际上都是收集研究资料的方法，而应用社会科学的发展，不仅在于调查技术，即收集资料技术的发展，还在于研究方法和分析技术的发展。近年来，无论是定性研究方法，还是定量研究方法都有了长足的发展。

首先，计算机技术的发展可谓突飞猛进，它对当今社会生活的各个方面产生了巨大的影响，在悄悄地改变着社会科学的研究风格和研究方式的同时，也大大提升了社会科学学者的研究能力。这种影响表现在研究过程的各个阶段，从理论建构（概念映射）、问卷设计（专业的问卷设计软件）、调查实施（计算机辅助访谈、计算机辅助电话访问系统、网络在线调查系统）、数据录入（光学标记识别软件）到数据分析（包括文本、声音、图像资料的处理），甚至延伸到写作发表阶段。这样的过程发生在如社会学、经济学、政治学、心理学、教育学中，促进了学科之间的相互借鉴和交叉融合，至少在研究方法上呈现这种趋势。随着计算机计算能力的大幅度提高，20世纪80年代之后，统计学领域内发生了一场“革命”，主要表现在对定类和定序变量的建模能力的大幅度提高上，以及与分布无关的统计分析模型的发展之上，特别是基于“Resampling”（包括Bootstrap、Jackknife、Monte Carlo模拟等）的建模技术*。同时，计算能力的提高还带动了基于神经网络、动态模拟、人工智能、生态进化等新兴的分析和预测模型的发展。这些进展都为定量社会科学研究提供了更

* 对于当前一些国内尚无公认译法的模型、软件等的名称，本“译丛”都只给出了英文的表述，以免造成称谓上的混乱。

多的可供选择的工具。

亚德瑞安·E·拉夫特里(Adrian E. Raftery)依据社会学家所处理的数据类型,将定量社会学在美国的发展划分为三个时代:第一代起始于上世纪40年代,交互表是其主要处理对象,研究重点是关联度和对数线性模型;第二代起始于上世纪60年代,主要处理单层次的调查数据,Lisrel类型的因果模型和事件史分析是其研究重点;第三代起始于上世纪80年代后期,开始处理诸如文本、空间、社会网络等非传统的数据类型,目前尚没有形成成熟的形态。拉夫特里的综述,虽然更强调定量社会学研究对统计学的贡献,但也大致勾勒出定量社会学在国外的发展脉络。

从分析模型的角度来看,定量分析在以下几个方向有了突破性发展:

1. 缺失值处理:由于社会生活的复杂性,社会调查数据常常出现缺失值,传统的处理方式是忽略这些缺失值,或者用均值替代。但现在则倾向于用多重插值法(multiple implation)或者其他基于模型的方法进行处理。这些技术的发展,不仅会增强我们对数据的处理能力,而且将改变我们设计问卷的方式。基于这些技术,我们在不增加被访者负担的前提下,大大增加调查问卷的内容:每个被访者只回答问卷的一部分,然后通过对缺失值的处理,获得他们对未回答部分的估值。

2. 非线性关系:线性假定是经典定量分析的一个常见假定,但在实际研究当中,线性假定只能被看作是对社会现实的一个逼近和简化。面对具体的研究数据,如果没有理论上的明确指引(不幸的是,我们常常没有中程理论的指引),我们是无法在线性模型和非线性模型之间做出取舍的。但MARS模型的出现,让我们可以从经验数据当中获得最为拟合的变量之间的函数关系,而不必预先做出线性假定。这样,理论思考和数据分析就可以实现一个互动的循环过程,定量分析就不单单是对理论和假设的简单证伪过程,而是理论思维一个重要组成部分。

3. 测量层次:20世纪六七十年代的统计模型,大多要求数据的测量层次在定距以上,如因素分析,但社会学的调查数据却大多为定类或定序数据。对应分析、Loglinear、Logit、Logistic Regression、潜类分析、Ordinal Regression、Normal Ogive Regression等统计模型的出现,大大提高了定量社会学处理定类和定序数据的能力。

4. 测量模型:基于文化、社会、心理和认知等方面的考虑,在社会学界仍有人对问卷调查在中国的效度提出质疑。抛弃"本土化"的文化执著,我们更应当关注的是问卷调查的项目反应理论(item response theory),即被访者回答问卷题器时的过程模型。这方面的进展主要表现在两个方面:一是分解测量量表的成分,如Rasch model、IRT分析、Mokken分析等,二是将测量模型与因果模型或其他分析模型结合在一起,明确把测量误差引入到分析当中,充分评估它们对分析结果的影响,如结构方程模型。

5. 潜变量模型:与测量模型相关联的另外一个发展方向是潜变量模型,例如,潜变量分层分析(latent class analysis)、潜变量结构分析(latent structure analysis)、潜变量赋值分析(latent budget analysis)等。"潜变量"这一概念表明,我们可以通过测量"显变量"来测量无法直接观察的理论概念,如权力、声望、地位等。这样,理论和现实之间,通过"潜变量"到"显变量"的映射(测量过程),就有了联接的桥梁。

6. 分析单元的层序性:在定量分析当中,我们常常强调要避免出现"生态谬

误”,即分析单元的层次和结论或推论的层次不一致。与其相关的方法论争论是“宏观和微观”的问题。随着多层次模型的出现,我们可以同时考察多个层次上的问题,我们可以把个人放在其家庭背景中,再把家庭放在社区的背景下,考察个人层次的变量对社区变量的效应,或者社区层次的变量对个体行为的具体影响。在定量分析模型当中,“宏观和微观”的联接获得了建模技术上的支持。在这个领域当中,还有一个方向也值得关注:分析宏观层次的数据,对微观层次进行推论。

7. 社会网络模型:区分“关系数据”和“属性数据”,是把分析重点从个体/群体等社会单元转移到这些社会单元之间关系的第一步,社会网络模型是目前发展较快的一个定量分析领域,其理论根基是结构主义。社会网络分析目前仍然具有较浓厚的“形态学”特征(基于图论的缘故),但却为我们理解社会关系在社会空间上的形态奠定了基础,通过计算机模拟和研究社会网络的历期数据,研究社会结构的“发生学”性质模型也处在萌芽状态当中。

8. 系统动力学:如果说社会网络模型是在社会空间上拓展定量社会学的研究手段,那么,社会过程在时间上和物理空间上的属性,则是事件史模型、事件数模型、历期分析、Cox 回归、时间序列分析、Cohort 分析、状态空间模型等模型的研究对象。在这个领域,计量经济学为定量社会学研究提供了许多有益的范例。

9. 预测模型:上述模型仍然是在分析主义的范式下。有些社会学的应用研究,更强调模型的预测精度,而不是模型的认知价值,例如,社会趋势的预测。由于计算能力的提高,神经网络、基因算法、人工智能、模式识别等数据挖掘技术有了长足发展,已经出现了许多拟合经验数据的预测模型,比较成功的应用出现在计量经济学领域(如对股市的预测)。

10. 计算机模拟:对于社会学应用研究而言,研究的对象具有历史性、规模大、变迁的过程不仅漫长且表现某种渐进性,且因社会隔离/社会伦理原因无法接近或有实验禁忌等,无法直接进行观察和研究,这时计算机模拟就成为一个可供选择的替代方案。计算机模拟主要有两个类型,一是基于计算机网络的模拟:每台微机作为一个代理,整个网络作为“社会”,实时演化,如法国的 Swarm 计划;二是基于概念模型的系统,在计算机时间上,按照既定规则运行,较有名的研究是罗马俱乐部的《增长的极限》,常见的软件有 Simul, Arena 等。自然科学家对此方向似乎比社会学家更有兴趣。

定性研究方法一直是社会学研究领域中比较传统的研究方法,在社会学研究的古典时期,它甚至是社会学家手中唯一的研究方法。但随着定量研究方法在社会学研究中的广泛应用,定性研究方法就似乎越来越不受到人们的重视。但需要澄清的事实是,在定量分析模型取得飞速发展的同时,在过去的二十多年里,定性研究方法也有了长足的进步。主要表现在以下六个方面:

1. 研究素材日益扩大:除了传统的参与观察、深度访谈、专题小组访谈之外,会话、交谈、电视、广播、文档、日记、叙事、自传(autobiography)等社会过程中自然产生的素材,甚至社会学理论本身(理论的形式化),也开始进入定性分析的视野当中。所有这些资料,不仅可以以文本的格式存储,而且,新型的多媒体介质,如图像、声音和视频,作为原始的分析素材,也日益成为定性分析的新宠。

2. 分析方法更加多样:定性方法的种类在最近的二十多年中,更是有了一个质

的飞跃。在比较传统的、源自语言学的方法，如内容分析、话语分析、修辞分析、语意分析、符号学、论据分析等方法之外，社会学家也创造出自己独特的定性分析方法，如施特劳斯(Strauss)等人的扎根理论、海斯(Heise)的事件结构分析、拉津(Ragin)的定性对比分析、Abbott 和 Hrycak 采用最优匹配技术的序列分析、亚贝儿(Abell)的形式叙事分析(formal narrative analysis)、鲍尔(Bauer)等人的语库建设、Attride-Stirling 等人的主题网络分析和神经网络技术应用的定性分析领域。所有这些方法的一个共同特征是，把定性研究向更加系统、更加精确、更加严格、更加形式化的方向推进。

3. 认识论基础更加多元化：现象学、释义学和本土方法论(ethnomethodology)的认识论，一直是定性分析的大本营，但近年来，实证主义也开始逐渐为定性分析所接纳，解释和阐释之间，由激烈的对立关系，逐渐演变为相互融合。

4. 研究过程更加客观规范：定性分析的一个主要问题在于阐释过程中不可避免的主观性。为了尽可能消除"解释者偏见"和主观选择性，定性分析开始遵循严格的程序模板或程序规则，并尝试引入定量分析中的"信度"、"效度"、"代表性"等概念，通过编码和对比，再加上传统的定性分析标准，如可解释性、透明性和一致性，使得定性研究的过程更加规范、阐释的结果更加客观，研究的结论更加可信。

5. 研究过程更加有效率：这主要应归功于大量计算机辅助定性数据分析(CAQDA)软件的涌现。从上个世纪 80 年代以来，定性分析过程的数字化和计算机化，已经是一个不可逆转的大趋势。这种发展趋势与定性研究者的理论取向无关，不管他们的理论立场是实证主义、符号互动论，还是本土方法论，大多数定性研究者都在自己的研究当中，开始采用计算机来辅助定性资料的分析过程。据不完全统计，目前已经有二十多种定性分析的软件，分别隶属于德国、英国、法国、美国等国家。其中，有一些软件是国外研究机构的科研成果，可以免费使用，但比较成熟的定性辅助系统大多是商业软件。这些定性分析的辅助系统，不仅使得研究者从处理大量文字材料的繁复劳动中解放出来，而且能够让研究者共享他们各自分析的细节，从而改变了定性研究的流程和研究集体之间的合作方式。同时，由于采用数据库结构，定性资料的管理也更加方便，这就为组织大型定性研究项目(包括多个研究地点、多个研究对象、历时的定性研究)提供了新的可能性。越来越多的定性研究人员开始走出他们的摇椅，坐到计算机屏幕前，湮没在访谈资料和故纸堆中的定性社会学家的形象已经一去不复返了。

6. 定性研究和定量研究的结合更加紧密：在定量分析方法的教材中，定性研究常常被看作是定量研究的前期准备工作，但定性研究者却持完全相反的观点，他们一般认为定性方法是自成一体的，可以完成从形成概念到检验假设的全部研究过程。在实际的应用研究中，定性方法和定量方法常常是交织在一起的，例如，克劳(Currall)等人在研究组织环境重要的群体过程时，通过内容分析把 5 年的参与观察资料量化，然后用统计分析来检验理论假定。格雷(Gray)和邓斯坦(Densten)在研究企业的控制能力时，利用潜变量模型把定性方法和定量方法有机结合在一起。雅各布斯(Jacobs)等人在研究比利时的家庭形态对配偶的家庭劳动分工影响时，首先用定量方法对纵向调查数据进行分析，从定量分析的结果中，又延伸出对核心概念的定性研究。这三个研究分别代表了定量和定性方法相互融合的三个方向：

①克劳等人的研究代表着定性方法的实践者试图将定性数据尽可能量化的取向，近年来涌现出的处理调查数据中开放题器的编码问题的工具软件(如 Words at, Smarttext 等,注意:它们都是由著名的统计软件公司出品的处理定性资料的软件),处理定性资料的传统内容分析软件(如 Nvivo、MaxQDA、Kwalitan 等)也开始提供将定性资料转换到常用统计软件的数据接口,这些工具上的革新将加快这种趋势的发展。②格雷和邓斯坦的工作代表了"方法论多元论"的取向,即在应用研究过程中,通过核心概念的测量模型,把定性研究和定量研究结合在一起。③雅各布斯等人的工作则代表了一部分定量研究者对过度形式化的定量方法的不满,并试图通过定性方法加以弥补。在定量研究领域中,对"模型设定"问题的关注,是定量方法重新试图返回定性研究这种取向的另外一种表现。

与社会调查技术和社会研究方法突飞猛进的现实相比,我国学术界在这些方面的论著的出版似乎显得有些迟缓。虽然已经翻译了美国的一小部分经典定量分析教材,如布莱洛克(Blalock)和巴比(Babie)的教材,也有自己编写的一些教材,如袁方等人的《社会研究原理和方法》、卢淑华的《社会统计学》等,此外,偏重软件操作的还有郭志刚的《社会统计分析方法——spss 软件应用》、郭志刚的《logistic 回归模型——方法与应用》、阮桂海的《spss for windows 高级应用教程》等。在《社会学研究》等专业杂志上,也常常有一些定量分析的应用研究,可是专门的方法和应用模型研究却没有,也没有专门的方法研究期刊。仅就定量研究方法的介绍而言,也存在一些缺陷,主要表现在:

1. 原理和操作脱节。
2. 过分依赖某些商业软件,不全面。
3. 与中国的实证研究相脱节。
4. 不能反映当前方法研究的最新进展。

与定量研究方法相比,由于各种原因,定性研究方法的引进和介绍都比较少。在福特基金会资助的方法高级研讨班上,曾讨论过一些定性研究方法。在定性方法研究方面也有少数专著,如袁方和王汉生 1997 年出版的教程,陈向明 2000 年出版的专著。但总体说来,我们对定性研究方法还停留在初步介绍的阶段,主要的介绍也局限在定性研究的研究设计和资料收集的阶段上,对定性分析方法的介绍则没有能够反映出当代定性方法的最新进展。特别是,在定性分析工具(定性分析软件)的引进和研究上,基本上还是一个空白。虽然不乏一些出色的定性研究报告,但从方法研究上讲,我们才刚刚起步。当然,我们同时还应当注意到,在历史学领域,我国对定性资料的鉴别、考据和分析,积累了大量的经验和知识,这也应当是定性方法研究的知识来源之一,应努力加以发扬光大。

令人欣慰的是,社会研究方法的引进和出版方面相对滞后的状况终于有所改观。重庆大学出版社的编辑,以独到的学术眼光,逆当前出版界唯利是图的不良选题风气,投入了大量的人力物力,组织出版"万卷方法"。自 2004 年至今,已引进社会科学研究方法方面的专著十余种,在我国社会科学界已经引起了一定的反响。然而,更为可贵的是,重庆大学出版社并未以已经取得的成绩而自满,而是再接再励,在原有"万卷方法"的基础上,进一步组织出版"万卷方法—社会科学研究方法

经典译丛”。按我们的设想,“译丛”应该是一个开放的体系,旨在跟踪社会科学研究方法发展的前沿,引进和介绍这一方面的经典著作和最新成果。

“译丛”第一批有《抽样调查设计导论》、《社会科学研究设计原理》、《社会科学研究测量原理》、《社会科学研究分析技术》、《问卷设计手册》、《回归分析法》、《数据再分析法》、《抽样调查设计导论》、《社会网络分析法》、《广义潜变量模型》、《定性变量数据分析》和《复杂调查设计和分析方法》(书名也许有变化)等十余种,几乎囊括了研究设计、测量和分析方法的所有领域,涵盖从基础的回归分析到最前沿的潜变量分析和多水平模型等各种分析方法。无论是社会科学各专业的本科生、研究生,还是社会科学研究的学者都将从中有所收获。

“译丛”由中国社会科学院社会学所社会调查和方法研究室的多位研究人员担纲,主译者都是在社会研究方法各个领域中具有相当造诣的教师和研究人员。“译丛”的译者不仅仅把翻译看作是一个“翻译”,而且也把它看作是一次再学习和再创新。

我们期待“译丛”的出版能对社会研究方法的研究、应用和教学有所推动。

沈崇麟　夏传玲

2006 年 12 月于中国社科院社会学所社会调查与方法研究室

作者前言

我们更新和再版1995年首版的《复杂调查设计与分析的实用方法》的主要目的是,对重点内容的突出和拓展、显著提高其实用性以及对读者反馈意见的采纳。举例而言,模型辅助性估计现在涵盖了一整章的组群的估算;处理非样本误差的章节全部重新改写;分析复杂调查方法包括了更加复杂成熟的估算技巧;扩展了个案研究的章节(此处的个案研究事实上是指一个调查分析的例子——译者注);讲解了调查过程中质量控制的实用方法;跨国教育调查个案例子的引入增强了国际比较的视角。我们相信,通过以上及其他扩展和加强,本书满足了更为广泛的读者群。

计算方法在本书首版过后发生了重要的变化。我们将这些技术材料放到了本书的扩展网页中。扩展网页旨在增加方法的实用性和提供更多的教学工具,例子和个案研究可以使用人机对话的方式,同时,还可以下载程序、实际数据和其他辅助性材料。对我们而言,这样可以更灵活地更新技术性材料。

我们非常感谢在本书的写作过程中一些机构给予的支持。特别地,我们希望提及基瓦斯开拉(Jyväskylä)大学教育研究所,芬兰交通和通讯部,芬兰国家公共卫生研究院,芬兰社会保险研究院,芬兰统计局以及基瓦斯开拉大学。首席统计分析师安特洛·马林(Antero Malin)提供了跨国教育调查个案研究的材料;高级顾问维尔皮·帕什廷能(Virpi Pastinen)提供了调查过程中质量控制的个案研究。我们非常感谢这些帮助。

卡尔-埃里克·桑德尔(Carl-Erik Särndal)教授对于书中几个部分的详细评论非常有益。朱阿·拉皮(Juha Lappi)博士为其中一部分提出了有用的评论。我们还要感谢统计学博士研究生维萨·基温内米(Vesa Kiviniemi)和硕士研究生安梯·帕萨能(Antti Pasanen)为网页建设所做的技术性工作,以及硕士研究生埃琳娜·尼基里(Elina Nykyri)协助校对和最后阶段的其他工作。我们感谢匿名评审对于我们改写第2版的提议的评论。最后的感谢要送给Wiley & Sons耐心而又灵活的工作人员。

目录

第1章 导 论

Introduction

总体梗概

本书的内容是关于抽样调查,它们在概念上可以分为两个大的类别。在描述性调查中,需要精准而又有效地估计某些——通常是很少的——总体特征。例如,在商业调查中,根据商业机构的样本来估算出不同职业类别的平均工资。抽样设计中的统计效率非常重要。对于效率而言,分层和其他使用诸如企业规模的辅助信息的方法,在抽样和估算的阶段也十分有益。虽然在估算过程中,经常用到假设总体和其他模型。但是,描述性调查中的推论,仅仅是针对一个固定的总体。另一方面,由于分析性调查有着多重目标,因而可以涵盖一系列的内容。在做一个分析调查的抽样设计时,能够在统计效率和费用效率之间找到一个整体平衡。举例而言,在一个以访谈为形式的调查中,可以设计一个多级抽样,对其中最后一级被抽中家庭户的所有成员都进行访谈。虽然,这样的分群降低了统计效率,但它却是资料收集中最为实际和经济的方法。费用效率可能很高,但在面对大量各类变量时,通过分层和利用其他辅助性信息的所得,对于统计效率而言,则无关紧要。在分析调查中,描述性目的也有可能显得重要。但是,通常的目标是各类人群间均值和比例的差别,或是对数和线性模型的系数,而非描述性调查中的固定人群的总数和均值。因此,与描述性调查相比,统计检验和建模在分析调查中的作用更显重要。

描述性调查和分析调查都可以很复杂。比如,它们使用多级分层的整群抽样设计。理解抽样的复杂性,对于在两类调查中得出可靠的估算和分析至关重要,特别是对于研究变量的群内相关的整群效应更是如此。这影响到方差估算以及检验与建模的过程。当使用非等概抽样抽取人群的各部分时,为了获取具有理想统计特征——诸如对抽样设计而言的非偏性和一致性——的估计值,应当使用相应的加权。同时,在描述性调查和分析调查中,也应当使用元素加权来调整无应答情形,以及使用变量缺损值的推算。

所以,这两种调查有许多相似之处,并且,在实际中没有真正的区别。一个以描述为主要目的的调查也可以具有分析调查的特征,反之亦然。但是,概

念上的区别是有益的。而这也正是本书组织材料的一个主要考量。

讲解的内容

为了实用性,讲授复杂调查设计和分析方法的书应当涵盖抽样、估算、检验和建模。在我们设计的调查程序中,首先考虑样本选择的原则和技巧。接下来,要检验未知总体参数的估计值,以及相应的标准误的估计值,以使得在实际中特定样本设计的估算可操作、可靠和有效。这些议题在描述性调查的框架下的本书的第一部分(第2,3章)中得到讲解。

本书的第二部分讲解与分析性调查相关的估算和分析(第5,7,8章)。对于复杂分析性调查,需要更为高级的技巧来估算方差。但,我们主要的关注点在于检验和建模。由于单变量表格、双变量表格以及多元分析(包括定类数据、对数及线性回归)在分析性调查中的重要性,我们也选择讲解这些内容。与描述性调查和分析性调查都相关的,处理非抽样性误差的技巧诸如再加权与推算则被放在本书的两个部分之间(第4章)。第6章讨论的组群估算虽主要属于描述性调查,但与两类调查均相关联。

为了展示各种方法,我们将充分讨论从卫生和社会科学研究的实际调查以及官方统计数据中抽取出的例子和个案研究。最后(第9章),其他例子涵盖了各种议题包括旅行调查、商业调查、社会经济调查以及教育调查。我们在例子和个案中一共使用了七个不同的调查数据。表1.1给出了这些数据的小结和一些技术信息。表中有三种调查数据。第2章到第4章使用了加总数据(1)(来源:官方统计数据)来展示描述性调查中的抽样和估算。真实的调查数据(2)(来源:芬兰国家公共卫生研究院)和(3)(来源:芬兰社会保险研究院)在第5章到第8章中作为范例用来演示复杂分析性调查中的组群估算、方差估算以及多变量建模。其他真实调查数据(4)到(7)(来源:芬兰交通与通讯部、芬兰统计署、芬兰社会保险研究院、OECD的PISA国际数据中心)被用在第9章的个案研究当中。

为了更好地利用本书的实用性目的,读者可以访问本书的扩展网页。在那里有更为详细的例子和个案研究的内容,也可以下载相关程序和数据以备进一步对话式的练习。

在第2章和第3章中,为了估算三种总体参数值,将讨论基本和高级的抽样技术,它们是简单随机抽样、系统抽样、概率对应规模抽样(PPS)、分层抽样、整群抽样。这些参数值是总和、比率和中位值。这些参数值的估算分别提供了线性、非线性和抗扰估计值的范例。我们将始终使用一个较小规模的固定的总体,重点放在推导出各个抽样技术所对应的近似的抽样权重。我们将额外比较各个估计值的相对效率(以标准误为基准),而总体结构中的信息将被逐步加以利用。使用这些辅助信息用于两个目的:抽样设计和特定抽样设计的参数值的估算。这些信息的使用在各个抽样技术中是有差异的。它们在

基本技术中使用较少,但在其他诸如分层和整群抽样等高级技术中变得更重要。在模型辅助估算的框架下,我们将讨论分层后估算、比率估算和回归估算。我们将展示,恰当使用辅助信息将极大地提高估算效率。总和、比率和中位值的统计特征,诸如偏差与一致性,也将通过蒙特卡洛模拟技术来加以检验。这一过程在扩展网页上有更多的内容。其中,各种抽样设计中估计值的变化都得到了讨论。

表 1.1 例子和个案中使用的真实调查数据

调查名称	主要抽样单位(PSU)的类型	调查数据中层级、类群和抽样元素的数目		
		层级	类群	抽样元素
	普查数据			
(1)1991 年省级人口数据(一个省)	自治市	2	8 个地区	32 个自治市
	调整后用于教育目的的真实调查数据			
(2)芬兰小型卫生调查(30 ~ 64 岁男性)	城市	24	48 个城市	2 699 人
(3)职业健康保健调查(10 人以上的单位)	工业单位	5	250 个单位	7 841 人
	个案研究中使用的真实调查数据			
(4)游客交通调查	个人	25	(个人调查)	11 711 人
(5)工资调查	商业单位	25	744 个公司	13 987 人
(6)健康保障调查(一个层级)	家庭	1	878 个家庭	2 071 人
(7)PISA2000 年调查(七个国家)	学校	7	1 388 个学校	32 101 人

在第 5 章中,我们将使用另外的(近似)方差估算方法来扩展第 2 章和第 3 章中的方差估算方法。我们使用子总体的均值与比例来展示分析性调查中常用的比率估计值。我们将使用线性化方法和包括平衡半样本、折刀方法以及脱靴方法的样本再使用技术。这些方法用来演示,一个从小型芬兰卫生调查截取的两级分层整群抽样设计。选择这一例子的原因是,这个样本是一个真实的容易处理的设计。检验和建模的过程将使用几个比率估计值方差和协方差的近似值。线性化方法在估算一致性的方差和协方差时,考虑到了各种抽样的复杂性,其中包括整群、分层以及加权。这些近似方法将被应用于卫生保障调查的抽样设计中。它的设计比前一个调查稍微复杂些。第 6 章讨论由地区或是类似标准建立的组群的总和估计值。我们将使用基于设计的模型辅助技术来演示职业健康保健调查。

第 7 章与第 8 章讨论复杂调查数据的分析。在检验单一变量和双变量表格的拟合度、同质性和独立性假设时,我们主要使用两种方法:一是沃尔德类检验统计量,二是拉奥-斯科特类修正。这些检验统计量的主要目的是校正整

群效应。这些检验的基础假设为,统计量是特定自由度的渐进性卡方分布;这一假设假定了大样本以及较大数量的样本整群。对于某些只有少数样本整群的设计,需要对检验统计量作出相应的自由度校正,导出了 F-分布的统计量。

我们在第 8 章转向多变量调查分析,其中有一个二分变量或是连续变量与一组预测变量。在使用对数和线性模型来分析定类变量时,需要运用通用加权最小二乘估算法。另外,在其中一些预测变量为连续性变量的对数和线性回归中,我们使用类似然和通用估算方程方法。为了恰当使用以上任一种方法,我们给出相应的分析选择方案的建议。在完全基于设计的方案中,由于充分考虑到了所有抽样的复杂性,因而是一个总体上有效的分析复杂调查的方法。在衡量加权、分层以及整群对于估算和检验结果的效应时,我们以基于简单随机抽样假设的分析方案作为参照对象。第 9 章的个案研究例子中,更多地使用了这些方案,来进一步演示多元分析。

本书中基于设计的分析的主要方法是,将整群效应当做估算与检验中的干扰的干扰(或是聚合)方法。这一方法的主要目的是,剔除这些干扰效应以获取有效的分析结果。在另外的分解方法中,整群效应本身有着内在的价值,也能提供有效的分析。我们在第 9 章最后一个例子中,使用适用于多层结构数据的多层模型来演示这一方法。

运 算

在设计一项不管是描述性或是分析性的调查时,都需要细心地规划所谓总体调查流程的各个环节。在通常情况下,一个调查流程以产生于实际信息需求中的提问环节开始。我们需要准备一个调查的总体计划,包括运用统计和调查方法的抽样、测量和分析设计。在调查的执行环节,需要评估和操作化整个计划。最后,要公开结果。在一个总体调查流程中,可以找出与本书相关的统计操作。图 1.1 列出了必要的方法和技术。

环节(1)中准备的计算机化的总体框架,是环节(2)抽取样本的基础。总体框架包括所有个体的辅助信息。这些辅助信息可以从多种渠道获得,例如人口普查和各种行政登记。这些数据在微观层次合并成一体(在实际中,这是可行的,例如使用在各个数据来源中个体唯一的身份标签)。收集来的数据在环节(3)得以清理。同时,选中的辅助资料也合并到数据中,以备估算和分析环节中使用。环节(4)纳入样本设计的标签到清理后的数据之中,以备环节(5)中的分析使用。因此,辅助信息可以在两个环节中用到:构建有效的抽样设计与使用模型辅助的估算技术来提高一个给定的样本效率。本书将广泛地讨论这两个环节。在实践中,环节(1)到(4)要使用用户特定的计算机程序。在环节(5)中,标准的估算和分析软件与用户特定的程序都可以使用。

为了实践中的方便,我们在例子和个案中使用可以购买到的软件作为我们在数据整理、调查估算和分析中的演示方法和计算工具。在本书扩展网页

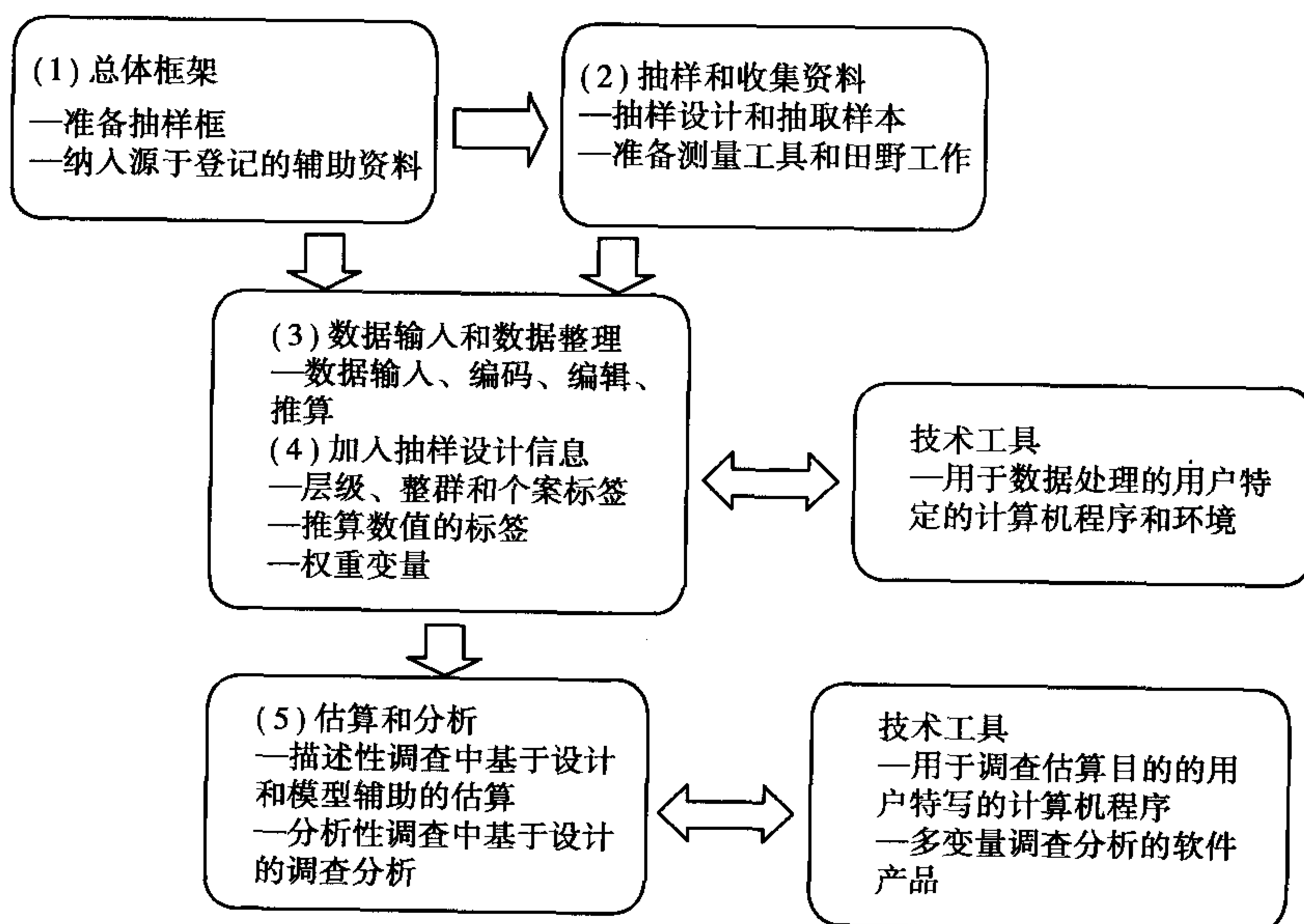


图 1.1 复杂调查中基于设计的估算和分析流程图

中，我们给出了技术性更强的方法和计算工具的讨论。

本书的使用

本书的读者群是研究者、抽样调查设计者以及工作在描述性或是分析性调查的计划、执行或是分析阶段的统计咨询师。我们旨在以简练的形式，为这些人员提供涵盖设计和分析复杂调查的最新方法的实用手册。通过使用带有计算指导和计算机化的范例的真实数据，本书也让读者对方法有更深的理解。因而，我们提倡读者参考本书的扩展网页。在网络上，许多实际例子被加以扩展并有更加详细的演示。那里还提供了一个进一步训练的方案，包括下载程序的可能性，以及在用户自己的计算机上进行对话式分析的实际数据。

本书的材料也可以用于大学层次的方法课程。第一个课程可以是从第 2 章到第 4 章的调查抽样，学生也可以从其中的小规模总体中学习实际抽样和估算。第 5 章到第 8 章提供了另一个更高级的课程。在两个课程中，本书的扩展网页可以用来支持教学和学习。同时，网页还提供了一些有用的数据。这些数据可用于练习复杂调查中的近似方差估计、假设检验、多变量分析。第 4 章也可以放入到更高级的课程中。第 6 章可以纳入组群估算的课程中。

第2章 基本抽样技术

Basic Sampling Techniques

在本章中，简单随机抽样、系统抽样和概率对应规模抽样被介绍为基本抽样技术。我们首先讨论抽样、抽样误差以及一个给定抽样计划的估算。同时，也给出一些关键概念的定义。

抽样和抽样误差

调查抽样考查的是一个固定的有限人群，其中每一个元素都被标注了独特的标签以便确认。概率抽样是一个从这样一个固定人群中选择随机样本——也简称样本——的灵活方法。概率样本的一个关键特征是每个元素均被赋予了一个正的被选中的概率；这一概率并非一定相等。选择样本时使用一个明确的抽样方案。抽样方案是指一系列选取样本的技术或规则。根据抽样方案的概率定义，样本的组成部分就被随机化了。

原则上，使用特定的抽样方案，可以从一个总体中抽取多个不同的样本。根据被选中的元素，可以从样本中得到不同的未知的总体参数值，例如总和——某一变量的取值的总和。抽样误差表示从潜在样本中计算得出的估计值的偏差。在一个特定调查的抽样设计过程中，抽样方案所产生的抽样误差越小越好。为了达到这一目标，对于总体结构的了解大有裨益。本章和第3章将讨论各种抽样情形下抽样方案与总体结构的关系。在这样的讨论中，无偏差估计值的标准误被当成抽样误差的测量值，设计效应将被用于各种抽样方案的抽样误差的比较。

样本的估算

使用特定的抽样方案选取一个样本之后，对于样本中选中元素，记录感兴趣变量——称为研究变量——的取值。资料收集上来后，可以进行统计分析。例如，我们经常会计算研究变量总体的总和以及它的标准误。在本章与下一章，我们讨论设计可行性较高的抽样过程的实用方法，以及在一个给定的抽样方案下恰当估算的实用方法。为了这一目的，首先让我们讨论考量抽样方案

在估算过程的作用的不同的方法。

实际分析调查的时候，经常要强调，估算要考虑抽样方案的结构。因此，分析中使用了所谓基于设计的方法。基于设计的方法的一个根本特性是，任何抽样方案的复杂性都将在估算过程中得以合理的考量。例如，这些复杂性可以产生于元素有着不同的被选中的概率。我们将在本章和第 3 章中进一步讨论这些。由于固定有限总体的元素有独特的标志，所以基于设计的方法中，这些抽样方案的特性可以结合到估算过程中。使用这样的标志，抽样设计的标签可以被纳入样本数据和分析之中。本章和下一章中，我们将详细地讨论各种抽样方案中怎样使用抽样标签。

本书使用不考虑抽样复杂性的分析作为基于设计的方法的参照。特别的，在比较更为复杂的抽样方案的效率时，放回式简单随机抽样——每个元素有着相等的被抽中的概率，每个抽中的元素又放回到总体中——有时被当成参照设计。在基于设计的方法中，假定有限总体为一个超总体的表现形式是非常有用的。与相应的辅助信息相结合，这一假设可以用于估算有限总体参数的模型构思中。当辅助信息已经被模型纳入估算过程，但推论仍然基于设计时，我们称之为基于设计的模型辅助方法，或简称为模型辅助方法(Särndal et al.,1992)。这一方法将在第 3 章中引入，在第 6 章中进一步得到讨论。

让我们进一步讨论基于设计的模型辅助方法，以演示模型假设可以用来简化特定抽样方案的估算过程。假定，轮船公司想要知道轮渡上所有旅客大约的总重量。这一信息对于未来做规划相当重要。测量所有旅客的体重太花费财力和时间，因此抽样在这样的情形下比较合适。假定，测量每十个旅客的体重，这样产生了一个有 n 个旅客的数据：$y_1,\cdots,y_k,\cdots,y_n$。研究者面临使用样本观测值来估计旅客的总重量的问题。同时，还要评估估计值的精确度。

在估算旅客的总重量时，研究者注意到，样本是从一个明确的有限总体中，依照特定的抽样方案抽取的。显然，使用了系统抽样。同时，从旅客登记中，上船的所有旅客的总数 N 提供了额外的信息。总重量的估算公式可以很容易地定义为 $\hat{t}=N\bar{y}$。其中，$\bar{y}=\sum_{k=1}^{n}y_k/n$ 是 n 个旅客的样本均值。为了估算抽样误差，$\hat{t}$ 的标准误应当被估计成方差估算公式 $\hat{v}(\hat{t})$ 的平方根。为了得到 $\hat{v}(\hat{t})$，研究者使用教科书上的方差估算公式 $\hat{v}_{srs}(\hat{t})=N^2(1-n/N)\hat{s}^2/n$。这一公式适用于无放回式简单随机抽样，其中的 $\hat{s}^2=\sum_{k=1}^{n}(y_k-\bar{y})^2/(n-1)$ 是旅客体重的样本方差。

使用上述公式得到的估计值通常已经能够满足实际目的了。但是，更进一步审视这一估算问题是有意义的。事实上，研究者在把 $\hat{t}$ 的方差当成简单随机抽样的估算公式来估算时，有一个简化程序的假设。而系统抽样的方差公式更恰当，因为应当纳入作为设计参数的组内相关系数 ρ_{int}。这两个方差估

算公式是相互关联的，有公式为 $\hat{v}_{sys}(\hat{t}) = \hat{v}_{srs}(\hat{t})[1+(n-1)\hat{\rho}_{int}]$。其中，$\hat{v}_{sys}$ 是系统抽样情形下的方差估算值。

但是，方差估算公式 $\hat{v}_{sys}(\hat{t})$ 并不适合实际。这是因为，每个抽样间隔中只抽取了单一元素。因此，如果没有关于旅客上船顺序的辅助信息，或是没有上船顺序的简化模型假设，就不能从已选样本中得到估算公式 $\hat{\rho}_{int}$。

最简单的模型假设是，旅客上船的顺序是完全随机的。这种情形下，组内相关为 0。因此，从系统抽样中估算出 $\hat{t}$ 的方差与从简单随机抽样中得出的方差相同。在这一简化模型假设的过程中，我们隐含地以超总体假设的形式，使用了基于设计的分析的辅助信息。系统抽样的使用辅助信息的替换方法，亦即模型假设，将在章节 2.4 中加以讨论。在那里，我们将演示，恰当地使用辅助信息，不仅可以简化估算过程，还更能有效地进行估算。

在本章和下一章中，我们介绍 5 种不同的抽样技术，并使用基于设计的方法估算一些总体参数和相应的标准误。对于每一个抽样方案而言，推导恰当的个体权重 w_k 至关重要。在上面的例子中，权重等于 N/n，亦即样本中元素被选中的概率的倒数。以上推导对于简单随机抽样和系统抽样均为有效。但是，对于更为复杂的抽样方案，所有元素的权重并不一定要相等。根据给定的抽样方案来推导估计值和估计值的标准误，以便正确的权重被纳入到公式中去。另外，也要指出在何种程度上以及以何种方式，特定抽样方案的总体辅助信息可以被使用。除了在抽样过程中使用这一信息外，它也可以用于意在降低标准误而选取的样本中使用模型辅助的方法的估算过程，以及获取接近于总体取值的估计值。这里，可以推导出称为 g 权重，记为 g_k 的新的权重。它的取值取决于抽取的样本以及选中的模型辅助估计值。

2.1 基本定义

现在，我们给出第 2 章到第 4 章的正式框架和基本定义。对应于各自的辅助信息，各种抽样方案也得以简要的描述。

总体和变量

一个有 N 个元素的有限总体 $\{u_1,\cdots,u_k,\cdots,u_N\}$ 的元素被从 1 到 N 标注。为了简便，用 k 来表示其中第 k 个元素。因而，这一有限总体可以写成：

$$U = \{1,\cdots,k,\cdots,N\}。$$

我们把未知总体取值的研究变量 y 写成 $Y_1,\cdots,Y_k,\cdots,Y_N$。在一些情形下，也用到额外的研究变量 x 和一个辅助变量 z。x 的未知总体取值写成 $X_1,\cdots,X_k,\cdots,X_N$。辅助变量 z 表示有限总体的额外信息，并通常假定各个元

素的取值已知。已知总体取值的辅助变量写成 $Z_1,\cdots,Z_k,\cdots,Z_N$。

总体参数

有限总体 U 的一个参数是研究变量 y 的总体取值 Y_k 的一个函数。某些情形下,这一方程还包含研究变量 x 的总体取值 X_k。典型的参数包括总和、比率和中位值。它们被定义为:

$$\text{总和 } T = \sum_{k=1}^{N} Y_k = Y_1 + Y_2 + \cdots + Y_N;$$

$$\text{比率 } R = \frac{T}{T_x}, \text{其中}, T_x \text{ 是研究变量 } x \text{ 的总体总和};$$

$$\text{中位值 } M = F^{-1}(0.5), \text{其中}, F \text{ 是 } y \text{ 的总体分布函数。}$$

选择总体总和的原因是它在调查抽样中的重要性。出版官方统计数字的统计机构在实施描述性调查时特别需要关注这一参数。许多抽样调查的经典文献都是有关总体总和估算的。因为总体均值$\overline{Y}$是总和的一个转换,即,$\overline{Y} = T/N$。后面出现的总和的估计值经过小转化也同样使用于均值。与均值不同,中位值的纳入是因为,正如后面演示数据所表现出来的,它恰当地表示了位置。比率是作为一个更复杂的参数引入的,它经常在实际中用到。第 5 章到第 9 章的调查分析中,将广泛用到比率类估计值。

抽样设计和样本

抽样调查的目的是从总体 U 中抽取的一个样本,然后从样本中估算未知的总体参数 T,R,或是 M。一个样本是 U 的一个部分。一共可以从总体中抽取许多不同的样本。我们用 S 来表示所有从 U 中可能抽取的规模为 $n(n<N)$的样本集合。事实上的样本写成 $s=\{1,\cdots,k,\cdots,n\}$。因而,$s$ 是所有可能的样本集合 S 中的一个。为了从 U 中抽取一个样本,需要使用特定的样本选择方案。在一个抽样方案中,可以表明样本 s 的抽取概率。这一概率被写成 $p(s)$。正式地,函数 $p(\cdot)$被称为抽样设计。按照实际的抽样方案抽取样本,而样本总和、样本比率以及样本中位值等随机数量则从样本中计算出来,而抽样设计则决定它们的统计特征(期望值和抽样误差)。虽然在文献中抽样方案与抽样设计的定义有所不同,但是在下面我们将交换使用这两个术语。在本书中,它们指我们从一个固定总体中抽取一个样本的方法。

在一个固定的抽样设计 $p(\cdot)$中,每一个元素均有一个选中概率,它用来表示元素进入样本中的概率。对于元素 k,其选中概率用 π_k 来表示。它也被称作一阶选中概率。在我们介绍各种抽样技术时,将用到这样的选中概率。

如果抽中的元素在每一次抽取后又放回到总体中,元素出现在样本 s 中的次数,就有可能超过一次。这样的抽样设计是放回式。相反的,在无放回式抽样中,元素只可能出现在样本 s 中一次。虽然在实际中通常使用无放回式

抽样方案,但在复杂抽样设计中,放回式假设简化了估算过程。显然,当总体数目巨大而样本规模显著小于总体时,放回式抽样与无放回式抽样的差别就显得不重要了。

对于样本 s 中的元素,我们测量研究变量 y。y 的 n 个样本取值用小写字母 $y_1,\cdots,y_k,\cdots,y_n$ 来表示。在诸如估算比率 R 的某些情形下,数据中也包括研究变量 x 的测量值 $x_k(k=1,\cdots,n)$。我们简要地假设没有测量误差。除了研究变量外,数据还应当包含抽样设计的某些信息,即诸如层级和组群标签以及权重等设计指标物。辅助变量 z(或是多个这样的变量)也经常出现在数据中。在下面抽样技术中将详细介绍这些变量。

估算公式

总体参数的估算公式指一个特定的计算公式或是算法程序,它是用来计算所选样本的样本统计量。我们寻求抽样设计下的无偏的或是一致的估算值,这样其期望值随着样本规模 n 的增加等于或是非常接近总体参数。我们将讨论以下三个估算公式:

总和 $\hat{t} = \sum_{k=1}^{n} w_k y_k$,其中,$w_k$ 是个体权重;

比率 $\hat{r} = \dfrac{\hat{t}}{\hat{t}_x}$,其中,$\hat{t}_x$ 是 x 的总和估计值;

中位值 $\hat{m} = \hat{F}^{-1}(0.5)$,其中,$\hat{F}$ 是 y 的估计分布函数。

运用估算公式观测到的样本中的数值被称为估计值。

抽样设计 $p(\cdot)$ 和估算公式的集合即是策略。在第 3 章最后部分讨论模型辅助估算时会用到这一概念。

估算值的方差

总体参数的估计值在不同的样本中各不相同。由抽样引起的这样的差异性显示了由一个特定样本作出的推论的不确定性。我们用估计值的方差 $V_{p(s)}$ 来衡量样本间的离异。由于 $V_{p(s)}$ 取决于抽样设计,它又被称为设计方差。它的取值可以使用合适的方差估算值从实际样本中得到。这一估计值用 $v_{p(s)}$ 来表示。方差估计值的平方根即是估计值的估算标准误。

严格地讲,设计方差仅适用于无偏估算公式;对于有偏估算公式而言,应该使用叫做平均方差(MSE)的更常见的抽样误差。MSE 可以由设计方差加上一个估计值偏差的平方来表示。这里的偏差是估计值的期望值与相应的参数的差异。总体而言,调查估算中,我们偏好无偏或是近似无偏的估计值。所以,使用设计方差也是合理的。这也适用于偏差随样本规模增加而减小的一致估计值。

设计效应

不同的抽样设计使用不同的总体参数估计值的设计方差。一个便利的评价抽样设计的方法是,比较设计方差的估计值与从(期望的)同等规模的参照抽样方案中计算出的设计方差。通常,放回式或是无放回式简单随机抽样被选为参照。例如,对于总和 T 的估计值 $\hat{t}$,设计效应是两个设计方差的比率,其简写形式为 DEFF,被定义为:

$$\mathrm{DEFF}_{p(s)}(\hat{t}) = \frac{V_{p(s)}(\hat{t})}{V_{srs}(\hat{t})},$$

其中,$p(\cdot)$是实际中的抽样设计。显然,获取 DEFF 需要两个设计方差的数值。但在实际中几乎没有现存的这样的数值。我们却可以计算得出。实际中,我们是使用样本数据中相应的方差来估算设计效应。设计效应的一个估算公式如下:

$$\mathrm{deff}_{p(s)}(\hat{t}) = \frac{\hat{v}_{p(s)}(\hat{t})}{\hat{v}_{srs}(\hat{t})}。$$

更普遍地,设计效应可以用策略$\{p(\cdot),\hat{t}^*\}$来定义。其中,$p(\cdot)$表示抽样设计,$\hat{t}^*$表示总和 T 的一个明确的估计值:

$$\mathrm{DEFF}_{p(s)}(\hat{t}^*) = \frac{V_{p(s)}(\hat{t}^*)}{V_{srs}(N\bar{y})},$$

其中,$\bar{y} = \sum_{k=1}^{n} y_k/n$ 是 y 的样本均值,$\hat{t}^*$是在方案 $p(s)$ 下 T 的基于设计或是模型辅助的估计值,$N\bar{y} = \hat{t}$ 是简单随机抽样下基于设计的估计值,$V_{p(s)}$ 与 V_{srs}是相应的方差。例如,$\hat{t}^*$ 可以是一个总和的回归估计值(参见章节 3.3)。

原则上讲,当 DEFF 为 1 时,一个抽样设计与简单随机抽样同等有效;当 DEFF 小于 1 时,比简单随机抽样更为有效;当 DEFF 大于 1 时,则没有简单随机抽样有效。我们将运用上述原则中的设计效应统计量来比较不同的抽样设计或是策略的效率。

抽样和估算中辅助信息的使用

抽样框——用来抽取样本的所有元素的登记名单,通常含有额外的关于元素的信息。辅助信息也可以从其他渠道获取,例如行政登记和官方统计。辅助信息有助于设计抽样方案以及在使用实际样本作估算时提高效率。要想有用,辅助信息应该与研究变量的离异相关。

在样本选取阶段中辅助信息的使用如下:

简单随机抽样(SRS)　样本的抽取没有使用总体的辅助信息。因此,在评估使用了辅助信息的更复杂的设计的收获以及估算的提高时,简单随机抽样方案(放回式或无放回式)提供了一个参照。

系统抽样(SYS) 使用了抽样框中元素的排列顺序的辅助信息。例如，当研究变量的取值随着排列顺序而增加时，系统抽样看起来就比简单随机抽样更为有效。设计方差估计值的另外设计参数——组内相关系数——显示了排列顺序与研究变量取值间的相关关系。

概率对应规模抽样(PPS) 辅助变量 z 表示总体中元素的规模。根据这一变量，各个元素有着不同的选中概率。抽样误差的大小取决于研究变量 y 与辅助变量 z 间的关系。

分层抽样(STR) 总体首先被分成互不交叉的次级部分，叫做层级。抽样在各个层级里独立进行。总的抽样误差是各层内抽样误差的总和。如果层级可以捕捉到研究变量的大部分离异，则分层抽样比简单随机抽样更为有效。

整群抽样(CLU) 假定总体被称为整群的自然组成部分所划分。从这些整群的总体中抽取一个整群样本。如果整群内部的同质性较高，则整群抽样的效率将低于简单随机抽样。组内相关系数是整群抽样中一个重要的设计参数。它测量整群内部的同质性。

在复杂调查中，可以用这五种抽样方法设计一个易执行的抽样方案。要么使用其中某一特定的方法，或是更普遍的各种方法的组合。除了简单随机抽样的所有方案中，元素的辅助信息都是必不可少的。对于已经抽取的样本，辅助信息可以用于估算过程中。普遍的用法是使用模型辅助的估算。辅助信息在估算过程中的使用如下：

后续分层 使用一个定类辅助变量，可以把抽取的样本划分成互不交叉的后续分层。而随后的估算就是建立在这一分层的基础上。辅助模型则是方差分析(ANOVA)类别。如果后续分层的内部同质性高，则可以提高效率。

比率估算 假定已知连续辅助变量 z 的总和，辅助模型是回归类别(没有截距)。如果研究变量 y 与辅助变量 z 相关，则可以提高效率。

回归估算(regression estimation) 与比率估算相似，假定已知辅助变量 z 的总体，辅助模型是回归类别(含有截距)。如果 y 与 z 相关，则可以提高效率。

所以，辅助信息可以用于设计抽样方案；而对于现存的样本，它能够提高估算效率。原则上，利用辅助信息可以降低标准误。因此，这样的资料是值得收集的。

更多的参考文献

本书的主要议题是基于设计的调查的估算和分析。特别是，估算和分析阶段利用抽样设计复杂性的方法。基什(Kish,1965)早期的教科书也讨论了同样的议题。他在其中使用了大量的实例来介绍设计效应。洛尔(Lohr,1999)也从实用的角度，使用从实际调查截取的例子，演示了基于设计、模型辅助和基于模型的估算方法。使用可靠的理论和数学框架，桑德尔等人

(Särndal et al. ,1992)的著作更偏重数理,他们讨论了抽样调查许多重要的领域。读者可以从本书扩展网页上找到更多参考文献。

2.2 1991 年省级人口

在实际抽样调查中,我们感兴趣的是规模受限的有限总体。正如将要在本书后面分析实际调查样本中所看到的,真实的总体普遍来讲数目巨大。在真实调查中,能看出抽样误差的来源及其决定估计值的特征并不容易。因此,我们选择了一个界定严格的问题和一个规模较小的总体,来演示各种不同的抽样方案及其对抽样误差的影响。例如,总和、比率与中位值参数可以从目标总体中精确地计算出来,并与从恰当的样本中得出的估算值相比较。这样,我们对整个目标总体有了了解。这一有限总体只有 32 个次级元素,其中要抽取数目固定为 8 的样本。很快,就可以发现这一演示与真实的大规模调查有着巨大的差异。但是,这一演示可以帮助理清重要的概念和问题。例如,怎样决定抽样分布,抽样设计怎样影响估计值以及它们的设计方差。

为了展示这些主要思想,从芬兰官方数据中截取了标题为"1991 年省级人口"的小数据。这一数据将在第 2 章到第 4 章中作为抽样框。芬兰有 14 个省,其中之一被选来作演示。到 1991 年 12 月 31 日止,这个省共有 32 个自治市和 254 584 个居民。表 2.1 提供了这一数据。

1991 年省级人口数据包含以它们在整个调查过程中的目的分类的 3 种信息。第一阶段是样本设计,在此需要诸如标记的标签变量,以及分清总体各部分的层级和整群。这里,总体的构成部分是自治市,它们名称和登记号码可以作为标签。其他两类信息定义研究变量和辅助变量。

在芬兰的官方统计中,自治市的列表是先城镇后乡村按字母顺序排列。这样,系统抽样技术可以使用这一自然顺序。随后,自治市总体也可以划分成互不交叉的称为层级的部分。另一个总体分类是相邻 4 个自治市组合而成一个整群。所以,整群的总数目为 8。标签变量 STR(层级)和 CLU(整群)分别对应城市与农村以及相邻 4 个自治市。

在下面的计算中,1991 年 11 月 30 日的总失业人口——简写为 UE91,是研究变量。技术上的程序如下:使用某种抽样技术,抽取含有 8 个自治市的固定规模样本;根据这一样本,计算一个基于设计的 UE91 的估计值,并使用设计效应统计量,研究其效率。对于模型辅助的估算以及概率对应规模抽样,从人口普查中选取了一个辅助变量(见表 2.1 的脚注)。它是家庭户数,简写为 HOU85。选择 HOU85 的原因是,它现存于人口登记之中,并与研究变量 UE91 高度相关。图 2.1 显示 UE91 的频次直方图。由于其分布偏斜,均值并不是恰当的显示位置的统计量。因而,选用中位值。

表 2.1 1991 年省级人口

序号	标 签	STR	CLU	% UE	UE91	LAB91	POP91	HOU85
	城市			**12.67**	**8 022**	**63 314**	**129 460**	**49 842**
1	Jyväskylä	1	1	12.20	4 123	33 786	67 200	26 881
2	Jämsä	1	2	11.07	666	6 016	12 907	4 663
3	Jämsänkoski	1	2	13.83	528	3 818	8 118	3 019
4	Keuruu	1	2	12.84	760	5 919	12 707	4 896
5	Saarijärvi	1	3	14.62	721	4 930	10 774	3 730
6	Suolahti	1	5	15.12	457	3 022	6 159	2 389
7	Äänekoski	1	3	13.17	767	5 823	11 595	4 264
	农村			**12.63**	**7 076**	**56 011**	**125 124**	**41 911**
8	Hankasalmi	2	5	15.07	391	2 594	6 080	2 179
9	Joutsa	2	6	9.38	194	2 069	4 594	1 823
10	Jyväskylän mlk.	2	7	11.82	1 623	13 727	29 349	9 230
11	Kannonkoski	2	4	18.64	153	821	1 919	726
12	Karstula	2	4	13.53	341	2 521	5 594	1 868
13	Kinnula	2	8	13.92	129	927	2 324	675
14	Kivijärvi	2	8	15.63	128	819	1 972	634
15	Konginkangas	2	3	21.04	142	675	1 636	556
16	Konnevesi	2	5	12.91	201	1 557	3 453	1 215
17	Korpilahti	2	1	11.15	239	2 144	5 181	1 793
18	Kuhmoinen	2	2	12.91	187	1 448	3 357	1 463
19	Kyyjärvi	2	4	11.31	94	831	1 977	672
20	Laukaa	2	5	12.11	874	7 218	16 042	4 952
21	Leivonmäki	2	6	10.65	61	573	1 370	545
22	Luhanka	2	6	10.34	54	522	1 153	435
23	Multia	2	7	11.24	119	1 059	2 375	925
24	Muurame	2	1	9.79	296	3 024	6 830	1 853
25	Petäjävesi	2	7	15.08	262	1 737	3 800	1 352
26	Pihtipudas	2	8	13.02	331	2 543	5 654	1 946
27	Pylkönmäki	2	4	17.98	98	545	1 266	473
28	Sumiainen	2	3	12.80	79	617	1 426	485
29	Säynätsalo	2	1	10.28	166	1 615	3 628	1 226
30	Toivakka	2	6	11.72	127	1 084	2 499	834
31	Uurainen	2	7	16.47	219	1 330	3 004	932
32	Viitasaari	2	8	14.16	568	4 011	8 641	3 119
	全省			**12.65**	**15 098**	**119 325**	**254 584**	**91 753**

注:表中的代码分别表示了中部芬兰省各自治市的一些人口指标,% UE—失业百分比,UE91—失业人口数,LAB91—劳动力,POP91—1991 年人口数,HOU85—1985 年家庭数。

来源:芬兰统计局:1985 年人口普查;芬兰统计局(1992):芬兰统计年鉴,第 87 卷;芬兰劳工部(1991):雇佣服务统计,1991 年 11 月 30 日。

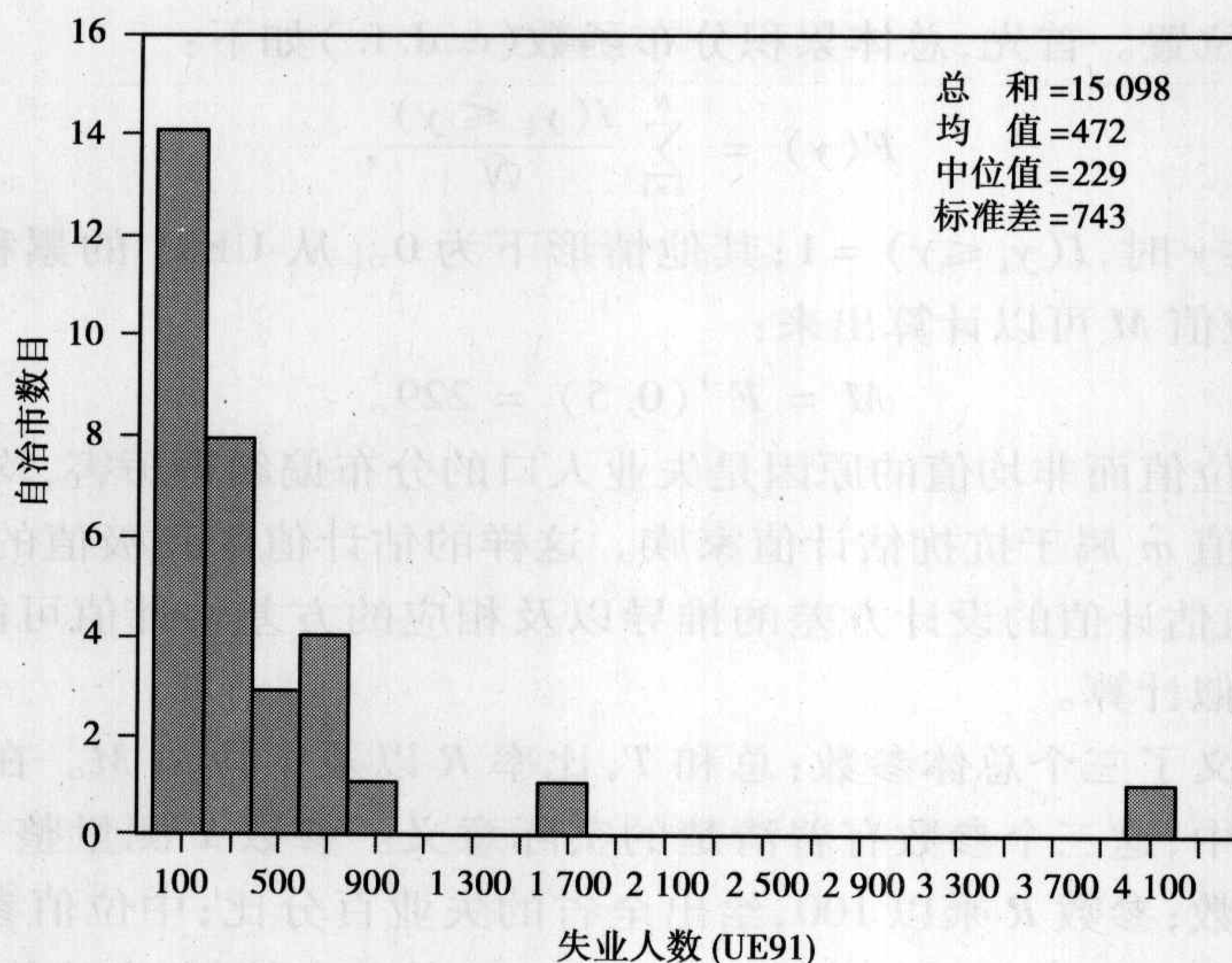


图 2.1 1991 年失业人数频次直方图(1991 年省级人口;$N=32$)

我们讨论 3 个总体参数:总和 T,比率 R 以及中位值 M。UE91 的总和是失业人口,总体总和如下:

$$T_{ue91} = \sum_{k=1}^{32} Y_k = 15\ 098。$$

另一个总体总和是劳动力总量 LAB91。它也能从表 2.1 中的数据中计算出来,其公式如下:

$$T_{lab91} = \sum_{k=1}^{32} X_k = 119\ 325。$$

最后,1991 年省级人口中的总人口数的数据为 254 584。在很久以来,经典抽样理论对总和都相当关注,而官方统计机构经常公布调查总体总和的估计值。

下面,T_{ue91} 仍然是各种抽样技术中估算的目标参数。它用一个数字来表示在该省中有多少失业人口的信息。由于估计值总和 $\hat{t}$ 是观测值的一个线性估算,它的设计方差以及相应的方差估计值相对简单和容易得出。

另一个有意义的总体参数是该省的失业率。它由两个总和得出:

$$R = \frac{T_{ue91}}{T_{lab91}} = \frac{15\ 098}{119\ 325} = 0.126\ 5。$$

这一比率的更实际的表示是失业百分比:$UE\% = 100R\% = 100 \times 0.1265 = 12.65\%$。

尽管参数 R 相对简单,在抽样设计并不复杂的情形下,比率的估计值 $\hat{r}$ 的设计方差也可能变得相当复杂。这是因为,比率估计值是非线性类型,需要在导出设计方差时进行近似计算。在经典抽样理论中,比率估算公式即是其估算的过程。我们将在章节 3.3 中讨论它。

第三个感兴趣的参数是中位值,或称为自治市失业人口在分布中的第 50

个百分位的位置。首先,总体累积分布函数(c. d. f.)如下:

$$F(y) = \sum_{k=1}^{N} \frac{I(y_k \leqslant y)}{N},$$

其中,当 $y_k \leqslant y$ 时,$I(y_k \leqslant y)=1$;其他情形下为0。从UE91的累积分布函数中,总体中位值 M 可以计算出来:

$$M = F^{-1}(0.5) = 229。$$

这里选择中位值而非均值的原因是失业人口的分布偏斜得厉害,均值 $\bar{Y}=472$。中位值估计值 $\hat{m}$ 属于抗扰估计值家族。这样的估计值受到极值的影响较小。但是,中位值估计值的设计方差的推导以及相应的方差估计值可能就相当麻烦并需要近似计算。

我们定义了三个总体参数:总和 T,比率 R 以及中位值 M。在1991年省级人口数据中,这三个参数有着清楚的实际意义。参数 T 测量整个省份里失业人口的总数;参数 R 乘以100,给出全省的失业百分比;中位值参数 M 提供了有关失业人口分布位置的信息。由于UE91的分布偏斜,它比均值更恰当。

在下面的例子中,我们将使用5种不同的抽样技术,从1991年省级人口数据中抽取含有8个元素的样本。它们是简单随机抽样(SRS)、系统抽样(SYS)、分层抽样(STR)、概率对应规模抽样(PPS)以及整群抽样(CLU)。抽样产生抽样误差。抽样误差根据抽样设计各不相同。但是,较小的演示总体的规模在计算上容易处理,这提供了分析抽样分布行为的机会。

2.3 简单随机抽样与设计效应

当没有任何关于总体结构的信息时,简单随机抽样可以被看成适用于这一情形的基本概率抽样方法。这一抽样技术保证每个元素有着相等的选中概率,因而能够得到一个可以很好代表总体的样本。

简单随机抽样有两个功能:一是它建立一个与其他抽样方法比较相对效率的基准;二是在诸如分层和整群抽样等更高级的抽样技术中,简单随机抽样可以成为抽取基本或是主要抽样单位的最终方法,也可以用于随机化。

本章节中的简单随机抽样显示,从总体中抽取一个部分作为样本总是引起计算中的抽样离异。正如1991年省级人口数据中全省的失业人口总和($T=15\ 098$),一个固定有限总体的参数是一个固定的常数。但是,如果从32个自治市中抽取一个含有8个自治市的样本,根据样本结构,失业人口总数的样本估计值 $\hat{t}$ 在不同样本中会各不相同。这样的离异将导致统计推断的不确定性,其生成的过程即是把它称为抽样误差的原因。

但是,在实际中,只分析一个样本。在统计推论中,抽样引起的离异应当得到控制。因而,研究者不得不熟悉未知总体参数的估计值的抽样分布。

下面,我们以三种抽样技术来介绍简单随机抽样:贝努利抽样(SRSBE)、放回式简单随机抽样(SRSWR)、无放回式简单随机抽样(SRSWOR)。通过从1991 年省级人口数据中以 SRSWOR 选择出一个含有 8 个自治市的样本,并进一步加以分析,这些抽样技术将得以演示。在这一基础上,将得出 3 个总体参数——总和 T,比率 R 以及中位值 M——的样本估计值。我们将使用调查估算软件来计算这些估计值,包括点估计、合适的标准误以及设计效应估计值。

最后,通过从 1991 年省级人口数据中模拟 1 000 个蒙特卡洛样本并计算样本分布的均值和方差,我们检视抽样误差的变动行为。在无偏估计的情形下,估计值抽样分布的均值应当等于被检视的参数,而模拟分布的方差应当接近于估计值的设计方差。像 1991 年省级人口数据这样的已知的固定总体的设计方差,可以被精确地计算出来。在讨论简单随机抽样的最后,我们给出设计效应参数和相应的从实际样本中得来的估计值。

抽取样本

简单随机抽样可以使用三种具体的抽取技术:贝努利抽样、放回式简单随机抽样和无放回式简单随机抽样。在第一种方法中,样本规模并不提前确定;在其他两种方法中则是确定的。贝努利和无放回式抽样中的样本抽取,可以很方便地使用应用于数据库的依次排列的过程。另一方面,放回式抽样中,每一次抽取都使用彩票法或是依次抽取的过程。所有这些技术均属于等概抽样设计(EPSN)的类型,其中的选中概率 $\pi_k = \pi$,即所有元素的概率均为同一常数。

贝努利抽样(SRSBE)　首先要预先确定选中概率。在这一情形下,常数 π 适用于所用元素,其中 $0 < \pi < 1$。常数 π 的取值是固定的,因而样本规模的期望值或均值是 $E(n_s) = N\pi$。在实际中,样本抽取时在抽样框加入两个变量:一个是取值相等或是取值为 π 的变量选中概率(PI);另一个是从均匀分布的(0,1)中取值的变量 EPSN。当 EPSN $< \pi$ 时,第 k 个元素就成为样本中的一员。使用这一程序,总体中的元素逐个得以选择。这一方法可以得到期望值为 $E(n_s) = N\pi$ 的样本规模,方差为 $V(n_s) = N(1-\pi)\pi$。这在小样本中产生一个方差估算的问题,但在大样本中样本规模的变化相对较不重要。注意,贝努利抽样是无放回式抽样的一种。

放回式简单随机抽样(SRSWR)　放回式简单随机抽样是在每一次抽取之后把所选的元素又重新放回到总体中的彩票似的选取。每个元素的选中概率在每次抽取后保持不变,而任意两次不同的抽取均是相互独立的。这样的特征也就解释了它是许多统计理论研究中的默认抽样技术。因为放回式的假设极大地简化了估算公式——特别是方差估算公式,在更复杂的抽样中,它经常被当成近似满足。SRSWR 设计也经常被用作设计效应计算中的参照设计。

无放回式简单随机抽样(SRSWOR)　在实际中最常使用的简单随机抽样

是无放回式简单随机抽样。为了简便，在公式中，SRS 用来代替 SRSWOR。单一元素的选中概率为一常数，但这与抽样进行到何种程度相关。这是因为每次抽取之后，留在总体中的元素的选中概率会增加。这就引起了计算方差估计值的困难，而前面的放回式抽样在这一方面要容易些。

表2.2 给出了一个从1991 年省级人口数据中可能抽取的规模为8 的 SRSWOR 样本。抽取的比率为 $n/N=0.25$。需要注意的是，这一样本可以由贝努利、放回式或是无放回式中的任一简单随机抽样获得。即使在复杂设计中，也可以假设，实际样本可以由这些基本抽样技术之一来完成。如果是这样的话，无放回式简单随机抽样也可以成为实际处理复杂抽样时计算设计效应时的参照。我们将使用刚刚抽取的样本作基于设计的估算。

表2.2 1991 年省级人口中一个无放回式简单随机样本($n=8$)

元素标签	研究变量	
	UE91	LAB91
Jyväskylä	4 123	33 786
Keuruu	760	5 919
Saarijärvi	721	4 930
Konginkangas	142	675
Kuhmoinen	187	1 448
Pihtipudas	331	2 543
Toivakka	127	1 084
Uurainen	219	1 330

抽样比例 =8/32 =0.25。

估 算

通过计算参数的点估计值和区间估计值以及进一步进行统计假设检验，统计推论从样本推及到目标总体。对于1991 年省级人口数据，感兴趣的关注点是总和 T，相对比例 $100R\%$ 及中位值 M，以及这些参数的点估计值及反映抽样误差的标准误估计值。在简单随机抽样的情形下，设计并不复杂，但在推导基于设计的估算公式、设计方差和这些方差的估算公式时，仍然可以突出基本特征。

从样本中计算相应的估计值时，也可以获得所期望的置信区间。同时，可以进行关于全省失业人口比例的统计假设检验。例如，我们可以检验这一百分比自去年以来是否保持不变，即，H_0: $100R\% = 100R_0\% = 9\%$ 。

让我们引入在无放回式简单随机抽样中，参数总和 T，比率 R 及中位值 M 的估算公式 $\hat{t}$,$\hat{r}$ 及 $\hat{m}$，以及相应的设计方差和标准误估算公式。对于总和 T,$\hat{t}$ 的标准公式是：

$$\hat{t} = N\bar{y} = N\sum_{k=1}^{n}\frac{y_k}{n} \tag{2.1}$$

或者是样本均值 $\bar{y}$ 乘以总体规模 N。估算公式也可以写成 $\hat{t} = \sum_{k=1}^{n} w_k y_k = (N/n)\sum_{k=1}^{n} y_k$，其中，$w_k = N/n$。常数 N/n 是抽样权重，也是抽样份额 n/N 的倒数。总和的估算公式也可以有另外的表达方式。首先，要定义总体中元素的选中概率。在 SRSWOR 情形下，元素 k 的选中概率为 $\pi_k = n/N$，并对于每一元素均为常数。在选中概率的基础上，总和的估算公式可以写成一个更普遍的霍维茨-汤普森类型的估算公式：

$$\hat{t}_{HT} = \sum_{k=1}^{n} w_k y_k = \sum_{k=1}^{n} \frac{1}{\pi_k} y_k = \frac{N}{n}\sum_{k=1}^{n} y_k。 \tag{2.2}$$

在这种情形下，估计值 $\hat{t}$ 与 $\hat{t}_{HT}$ 显然相等。这是因为选中概率 $\pi_k = n/N$ 对于任一元素 k 均相等。霍维茨-汤普森类型的估计值，也经常用于选中概率并不固定的概率对应规模的抽样之中。对于抽样设计而言，这一估计值有着无偏的统计特性。

比率 R 的估计值是两个总和的比率：

$$\hat{r} = \frac{\hat{t}}{\hat{t}_x}, \tag{2.3}$$

其中，$\hat{t}_x$ 表示研究变量 x 的总和。虽然这两个总和的估计值均是无偏的，但比率估计值 $\hat{r}$ 还是属于有偏估计值的类型。让我们更进一步讨论 $\hat{r}$ 的偏差。

$\hat{r}$ 的偏差与两个变量 y 和 x 之间的线性回归有关，它的形式是 $y = A + Bx$。当截距 $A = 0$，回归直线经过原点，这意味着比率 Y_k/X_k 对于所有总体中的元素为一常数。这时，比率估计值 $\hat{r}$ 是无偏的。但是，当 $A > 0$ 时，偏差等于：

$$\mathrm{BIAS}(\hat{r}) = E(\hat{r}) - R \approx V_{srs}(\bar{y})\frac{A}{\bar{Y}^2\bar{X}}, \tag{2.4}$$

其中，$V_{srs}(\bar{y})$ 表示在 SRSWOR 设计中 $\bar{y}$ 的设计方差，$\bar{Y}$ 和 $\bar{X}$ 是研究变量 y 与 x 的总体均值。

这一公式表明，当常数 A 较大时，偏差也相当大。另一方面，当样本规模增加时，方差 $V_{srs}(\bar{y})$ 变小，这将导致偏差的降低。所以，$\hat{r}$ 是 R 的一致性的估计值，并随着样本规模的增加而被认为更加可靠（图 2.3 显示了有限总体的一致性）。

为得到中位值 M 的估计值，首先需要求得研究变量在点 y 的累积分布。霍维茨-汤普森类型的累积分布函数（c. d. f.）的估算公式为：

$$\hat{F}(y) = \sum_{k=1}^{n} \frac{w_k I(y_k \leq y)}{\hat{N}}, \tag{2.5}$$

其中，w_k 表示第 k 个样本元素的权重。当 $y_k \leq y$ 时，$I(y_k \leq y) = 1$；其他情形下为 0。权重的总和是 $\hat{N} = \sum_{k=1}^{n} w_k$。这一估计的 c. d. f. 是一个非连续性函数。为求取中位值 M 的估计值 $\hat{m}$，首先应当平滑化这一函数。平滑化分布函数的方法是用直线连接 $\hat{F}(y)$ 各点。从中可估计四分位值——也包含了中位值。这

一程序给出了中位值的无偏估计量。佛朗西斯科和富勒(Francisco and Fuller, 1991)有更多的细节。

为了确定置信区间和检验统计量,需要 $\hat{t}$,$\hat{r}$ 及 $\hat{m}$ 估计值的设计方差或是这些方差的估计值。它们用来估计由从总体中随机抽取样本而产生的标准误。在这里,我们推导适合单一样本情形的方差估算公式。更为普遍的抽样误差的行为变动,将在估计值的设计方差和抽样分布中另行处理。

总和的估计值 $\hat{t}$ 的设计方差 $V_{srs}(\hat{t})$(见式 2.8)的无偏估计如下:

$$\hat{v}_{srs}(\hat{t}) = N^2\left(1 - \frac{n}{N}\right)\sum_{k=1}^{n}\frac{(y_k - \bar{y})^2}{n(n-1)} = \frac{N^2\left(1 - \frac{n}{N}\right)\hat{s}^2}{n}, \qquad (2.6)$$

其中,$\bar{y} = \sum_{k=1}^{n} y_k/n$ 是样本均值,$\hat{s}^2 = \sum_{k=1}^{n}(y_k - \bar{y})^2/(n-1)$ 是元素方差 S^2 的估计值。方差估计值的平方根是估计值 $\hat{t}$ 的标准误,写成 s. e($\hat{t}$)。

由于比率 $\hat{r}$ 和中位值 $\hat{m}$ 的方差估算公式应该被当成非线性的估算,因而它们的估算更加复杂。比率估计值 $\hat{r}$ 的近似方差估计值是:

$$\hat{v}_{srs}(\hat{r}) = \left(1 - \frac{n}{N}\right)\left(\frac{1}{\bar{x}^2}\right)\sum_{k=1}^{n}\frac{(y_k - \hat{r}x_k)^2}{n(n-1)}。 \qquad (2.7)$$

在推导方差估算公式时,使用了泰勒级数展开式来线性化比率估算公式。因此,上述等式给出了设计方差的近似估计值。第 5 章将讨论更多的细节。$\hat{m}$ 的方差估算公式也需要利用线性化的方法。这意味着,特别是对于小样本来讲,中位值的估算公式并不十分稳定。中位值的标准误确定如下。首先,建立一个低限为 0.975 水平和高限为 0.025 水平的平滑的累积分布函数。第 p 个四分位的标准误,是这一平滑分布函数的上限和下限之间在 p 点的水平距离的四分之一。

基于设计的估计值的计算

在这里和本书的其他地方,我们使用合适的、考虑了设计复杂性的软件来计算基于设计的估计值及其标准误。统计分析依循图 1.1 里的流程。我们假定数据是经过整理的,数据处理的过程是成功的。这一过程包括数据录入、编码、编辑、推算和抽样权重的推导。

在基于设计的估算中,以下分析所需的抽样设计的标签必须包含在数据中:层级标志变量、整群标志变量和抽样权重变量。应当注意到,除了复杂设计以外,这些标签还可以用于简单设计中——例如,只含有一个层级或是没有整群分类(每个单位本身就是一个整群)的设计中。除了这些变量,在无放回式抽样中,抽取的比率也是必需的。用户熟悉的计算机程序经常用于清理稍后用于分析的数据。

在分析阶段,抽样标签由调查分析软件所读取。当然,对设计信息的使用

需要充分了解到抽样设计的复杂性。我们将运用所有本书讨论的抽样技术来演示设计信息在估算过程中的使用。标准估算软件的结果输出包括点估计及其估计的标准误、离异系数以及设计效应。另外，某些有用的抽样设计的信息也包含在其中。我们第一个例子是无放回式简单随机抽样情形下的基于设计的估算。

范例 2.1

我们分析 1991 年省级人口数据的一个 SRSWOR 样本。从先前以无放回式简单随机抽样抽取的样本中，我们将得到总体、比率和中位值，以及它们的标准误的估计值。首先，把设计标签加入抽取的数据中。这包括分层标签（STR）。在简单随机抽样中，它是一个常数，即 STR = 1。其次，我们需要知道，某一元素是否属于一个组别或是整群。在元素抽样中，每一个元素其本身即是一个整群。因此，CLU 等于观测中的识别码（ID）。最后，我们输入权重变量。在简单随机抽样中，它等于选中概率的倒数，$w_k = \pi_k^{-1} = (n/N)^{-1} = N/n$。它在估算总和时，用来加权样本观测个案，所以权重的总和为 N。总体而言，在估算总和时，权重变量应当调整比例，以便权重的总和等于总体规模。在这一例子中，总体规模为 32 个自治市（$N = 32$），抽取的样本包含 8 个自治市（$n = 8$）。所以，权重变量的取值为 WEIGHT = 32/8 = 4。

当这些初步处理完成之后，数据应当与表 2.3 类似。为了便于阅读，我们加入了按字母排列顺序的元素标签 LABEL。其他变量分别列入两个栏目中："抽样标签"和"研究变量"。

表 2.3　给出样本设计标识的 1991 年省级人口的一个 SRSWOR 样本

样本设计标识			元素标签	研究变量	
STR	CLU	WGHT		UE91	LAB91
1	1	4	Jyväskylä	4 123	33 786
1	4	4	Keuruu	760	5 919
1	5	4	Saarijärvi	721	4 930
1	15	4	Konginkangas	142	675
1	18	4	Kuhmoinen	187	1 448
1	26	4	Pihtipudas	331	2 543
1	30	4	Toivakka	127	1 084
1	31	4	Uurainen	219	1 330

抽样比例 = n/N = 8/32 = 0.25。

在处理无放回式抽样中小规模总体的方差估算公式时，给出能够表示有限总体校正（f. p. c.）的抽样比例很重要。在这个例子中，抽样比例为 8/32 = 0.25。因而，有限总体校正等于 $(1 - n/N) = 0.75$。

表 2.4 给出了估算结果。其中有 $\hat{t}$，$\hat{r}$ 与 $\hat{m}$ 的点估计及其标准误、离异系

数与设计效应(deff)。总体的离异系数是 $c.v(\hat{t}) = s.e(\hat{t})/\hat{t}$。由于 SRSWOR 设计是参照方案,这里,deff 的估计值是单位元素 1。除了这些估计值,表中还有相应的总体参数 T,R 和 M 的数值。本书的扩展网页有更进一步的细节。

表 2.4 1991 年省级人口中一个 SRSWOR 样本估计值($n=8$)

统计量	变 量	参 数	估计值	s.e	c.v	deff
总和	UE91	15 098	26 440	13 282	0.50	1.00
比率	UE91,LAB91	12.65%	12.78%	0.41%	0.03	1.00
中位值	UE91	229	226	150	0.66	1.00

对这些估算的理解如下。全省总失业人口数(UE91)T 的点估计为 $\hat{t}$ = 26 440,相应的标准误的估计值是 $s.e(\hat{t}) = 13\ 282$。在这两个估计值的基础上,假定标准正态分布 $N(0,1)$是估算的总和的近似分布,可以得出全省失业人口总数的 95% 置信区间是:

$$\hat{t} - 1.96 \times s.e(\hat{t}) < T < \hat{t} + 1.96 \times s.e(\hat{t})$$

即,$407 < T < 52\ 472$。因为范围太宽,它几乎没有任何行政参考意义。稍后,我们将讨论选择产生更小标准误而更有效的抽样方案是怎样影响置信区间的。

该省失业人口百分比的估计值 $\hat{r}$ 为 12.78%。由于 $\hat{r}$ 的标准误估计值 s.e 已知,我们可以从统计上检验当前的失业率 R 是否与一年前的 9% 有所不同。因而,有 $H_0: R = R_0 = 0.09$。再次使用正态近似,我们有:

$$Z = \frac{\hat{r} - R_0}{s.e(\hat{r})} = \frac{0.127\ 8 - 0.09}{0.004\ 1} = 9.22^{***},$$

我们拒绝原假设,并得出结论,在上一年中,该省的失业百分比已经显著地变化了。显著水平用 * * * 来表示拒绝概率——假设检验的 p-值。在这里,其值小于 0.001。

另一方面,比率和中位值的点估计与相应的参数相当接近。

接下来,我们更详细地解析 $\hat{t}$,$\hat{r}$ 与 $\hat{m}$ 估计值的设计方差和抽样分布。

设计方差与抽样分布

在某种抽样设计的情形下,用简单随机抽样很容易演示不同的估计值及其方差是怎样变动的,以及随机化是怎样影响抽样误差的。在检视这一变动时,我们首先计算 $\hat{t}$,$\hat{r}$ 与 $\hat{m}$ 在 SRSWOR 设计中的设计方差——用 V_{srs} 来表示。从分析中的小规模固定有限总体中可以计算出这些方差。但是,这些设计方差并没有包含抽样误差的所有信息;这些估计值的抽样分布的推导过程,可以让我们更进一步地检视它们的变动情况。

得到这些抽样分布的通常的方法是,使用一个给定的抽样方案从总体中模拟抽取大量的样本。使用 SRSWOR,我们运用蒙特卡洛方法从 1991 年省级

人口数据中,模拟抽取了规模为 8 的共计 1 000 个样本。对于每一个样本,均计算了 $\hat{t}$,$\hat{r}$ 与 $\hat{m}$。总和、比率与中位值的估计值的分布即是该估计值的实验抽样分布。这些分布提供了抽样分布的位置和形状的相关信息。

在 1991 年省级人口数据的 SRSWOR 方法下,总和、比率与中位值的设计方差的公式及所对应的观测值如下。

总和 T:$\hat{t}$ 的设计方差是,

$$V_{srs}(\hat{t}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\sum_{k=1}^{N}\frac{(Y_k-\bar{Y})^2}{(N-1)} = \frac{N^2\left(1-\frac{n}{N}\right)S^2}{n}, \tag{2.8}$$

其中,$\bar{Y}=\sum_{k=1}^{N}Y_k/N$是总体均值,$S^2=\sum_{k=1}^{N}(Y_k-\bar{Y})^2/(N-1)$ 是总体方差。观测到的设计方差是,

$$V_{srs}(\hat{t}) = \frac{32^2}{8}\times\left(1-\frac{8}{32}\right)\times 743.36^2 = 7\ 283^2。$$

比率 R:$\hat{r}$ 的近似设计方差是,

$$V_{srs}(\hat{r}) \approx \frac{1}{\bar{X}^2}\frac{1}{n}\left(1-\frac{n}{N}\right)\sum_{k=1}^{N}\frac{(Y_k-R\times X_k)^2}{(N-1)}, \tag{2.9}$$

根据它得到的观测值是,

$$V_{srs}(\hat{r}) = \frac{1}{3\ 729^2}\times\frac{1}{8}\times\left(1-\frac{8}{32}\right)\times\frac{315.91^2}{(32-1)} = 0.005^2。$$

中位值 M:中位值 $\hat{m}$ 的设计方差有几个近似方差。一个可能是从累积分布函数中近似估算方差:

$$V_{srs}[\hat{F}(\hat{m})] = \frac{N-n}{N-1}\frac{1}{n}F(M)[1-F(M)] \approx \frac{1-\frac{n}{N}}{n}\times 0.25, \tag{2.10}$$

它非常简单,因为其中没有未知。它得出,

$$V_{srs}[\hat{F}(\hat{m})] \approx \frac{1-0.25}{8}\times 0.25 = 0.02,$$

为得到在普通研究变量计量单位上的 $\hat{m}$ 的设计方差,这一估计值应当重新调整计量单位。在 1991 年省级人口数据中,我们使用蒙特卡洛模拟(参见图 2.2)得来的近似设计方差。因而,我们有,

$$V_{srs}(\hat{m}) \approx \hat{v}(\hat{m}_{mc}) = 107^2。$$

注意,设计方差写成了标准误的平方的形式,是为了便于与表 2.4 中的标准误相比较。当比较某一估计值的设计方差或标准误与实际样本中得来的相应的估计值时,可以发现样本间的差异导致了不同。例如,总和的方差估计值是 $\hat{v}_{srs}(\hat{t})=13\ 282^2$,而计算出的相应的设计方差是 $V_{srs}(\hat{t})=7\ 283^2$。这里的样本估计值过分夸大了设计方差。比率估计值中,这些数值为 $\hat{v}_{srs}(\hat{r})=0.004^2$ 和 $V_{srs}(\hat{r})=0.005^2$,它们相当接近。最后,中位值的数值是 $\hat{v}_{srs}(\hat{m})=150^2$ 和

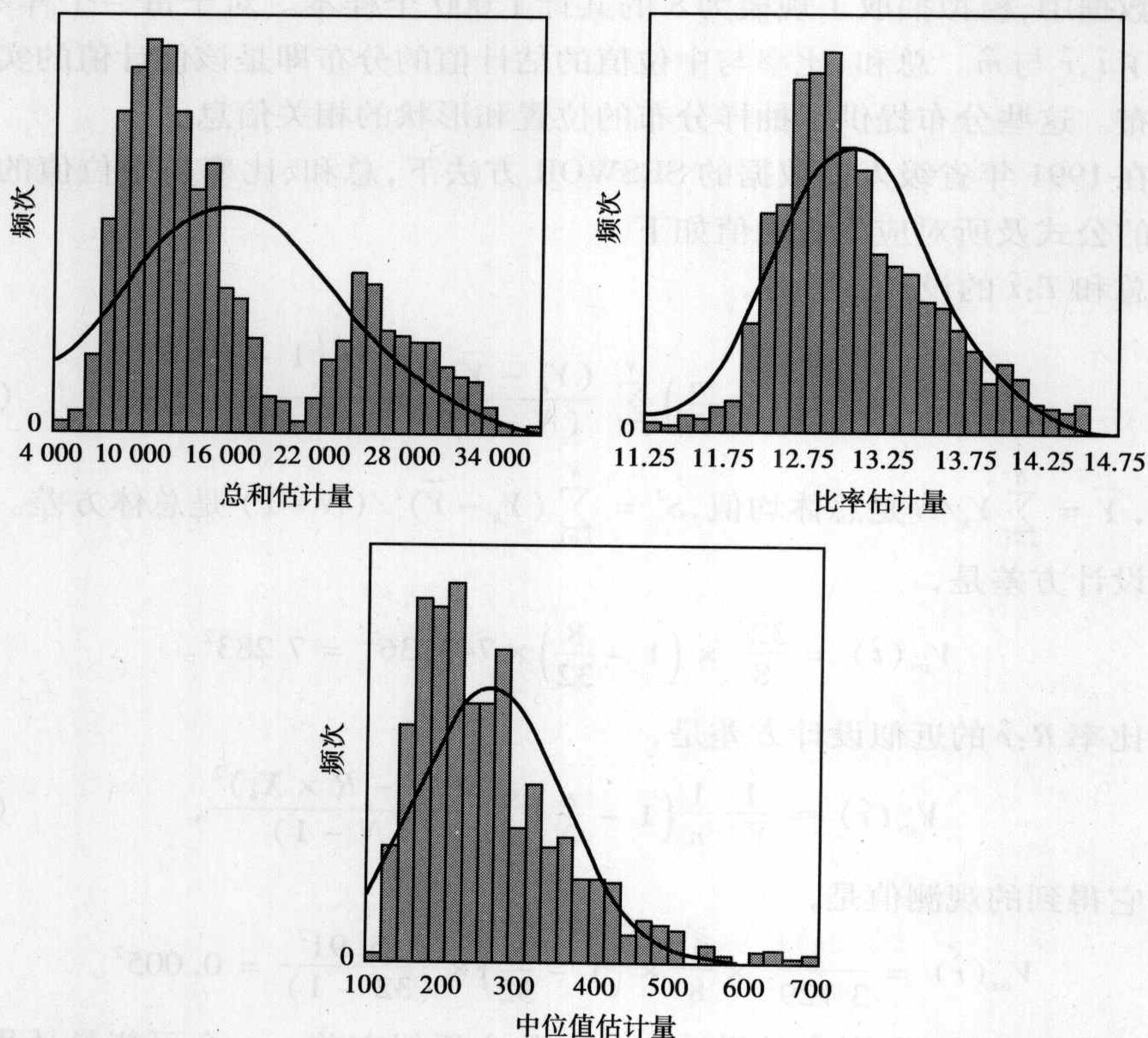


图 2.2 据 1991 年省级人口的 SRSWOR($N=32, n=8$)设计下

1 000 个蒙特卡洛样本估计量 $\hat{t}$,$\hat{r}$ 与 $\hat{m}$ 的抽样分布

$V_{srs}(\hat{m})=107^2$。样本估计值同样地比相应的设计方差大很多。

为了更进一步审视无放回式简单随机抽样条件下估计值的变动情况,蒙特卡洛模拟所得的总和、比率与中位值估计值在图 2.2 中用直方图表示了出来。蒙特卡洛估计值分布的均值的期望值与总体参数相等,其方差应接近于设计方差。

总和估计值的均值 $\hat{t}_{mc}=15\ 049$,与相应的参数 $T=15\ 098$ 符合;总和估计值的方差为 $7\ 278^2$,与设计方差 $V_{srs}(\hat{t})=7\ 283^2$ 相近。估计值 $\hat{t}$ 在这一方面也是合适的。

更进一步的审视,发现直方图有两个峰值。其分布与正态分布的钟形图案并不相像。正态分布可以用作参照(在图中,相应的正态分布的取值用实线来显示)。观测分布与理论分布有着相当大的差异。这让我们警惕使用正态分布的假设做推论。原因很简单。$\hat{t}$ 的抽样分布在很大程度上取决于 1991 年省级人口数据中 UE91 的分布。而 UE91 的分布极大地向共有全省三分之一人口居住的省府偏斜的(见图 2.1)。总体和样本的规模并不足以达到近似

正态的要求。相应的,对于此总体,简单随机抽样可能并不是估计总和的合适的技术。

模拟分布显示,对比率 UE91/LAB91 的估计 $\hat{r}$ 比较圆满。比率估计值的均值为 $\hat{r}_{mc}=0.128$,几乎等同于总体参数 $R=0.1265$。比率估计值的方差为 0.006^2,与设计方差 $V_{srs}(\hat{r})=0.005^2$ 相当。另外,其分布也几乎是钟形的,显示近似正态的假设比总和估算中要合适得多。

中位值 M 的定义是研究变量 y 的累积分布函数的第 50 个百分位取值。通常,c. d. f. 是未知的,应当使用近似分布。估算中位值的通常做法是,将样本中的取值 $y_{(1)}<\cdots<y_{(k)}<\cdots<y_{(n)}$ 从小到大排列。如果样本规模是奇数,则中间取值即为中位值;若不是奇数,则中位值是两个中间取值的均值 $\hat{m}=\frac{1}{2}[y_{(n/2)}+y_{(n/2+1)}]$。这种中位值的估算值经常被称为 50% 裁剪均值。

对于一个对称的总体来讲,均值和总体相同。正如图 2.1 所示,1991 年省级人口数据是极大地偏斜的。因而,总体均值与中位值的差异较大,为 $\overline{Y}-M=472-229=243$。接下来,我们考察样本规模对于总和与比率估计值变动行为的影响。

有限总体一致性与样本规模

现在,我们使用模拟的方法来检视两个基本估计值 $\hat{t}$(表示总和)与 $\hat{r}$(表示比率)的统计特征。

当从所有可能抽取的样本中得到的估计值与真实的总体取值完全相等时,这一估算方法被称为无偏。在 $n=N$ 或是样本包括整个总体情形下,当样本估计值与真实的总体取值完全相等时,这一估算方法被称为一致的(Cochran,1977,第 21-22 页)。在桑德尔等(Särndal et al. 1992,第 168 页)的书中,这一类型被定义为有限总体一致性。我们还是使用蒙特卡洛模拟方法,从 1991 年省级人口数据中抽取 1 000 个 SRSWOR 样本,来考察总和与比率估计值的变动情况。同时,我们使用不同的样本规模:从 $n=1$ 到总体规模 $N=32$。结果呈现在图 2.3 中。

如图 2.3(a)所示,研究变量 UE91(失业人口数)总数 T 的估计值 $\hat{t}=N\times\sum_{k=1}^{n}y_k/n$ 是无偏的。可以从图 2.3(b)中看出,其标准误 s. e($\hat{t}$)随着样本规模的增加而减少。另一方面,比率的估计值 $\hat{r}=\sum_{k=1}^{n}y_k/\sum_{k=1}^{n}x_k$ 在一定程度上是与总体比率 R(=0.126 5)有偏差的,但它却是一致的(图 2.3(c))。这里,x 是指研究变量 LAB91(劳动人口数)。一致性的证据是随着样本规模增加而偏差减少。同样的,比率的估计值的标准误也随着样本规模增加而减少(图 2.3(d))。我们得出结论,两个估计值都是一致的;同时,总和的估计值也是无

偏的。

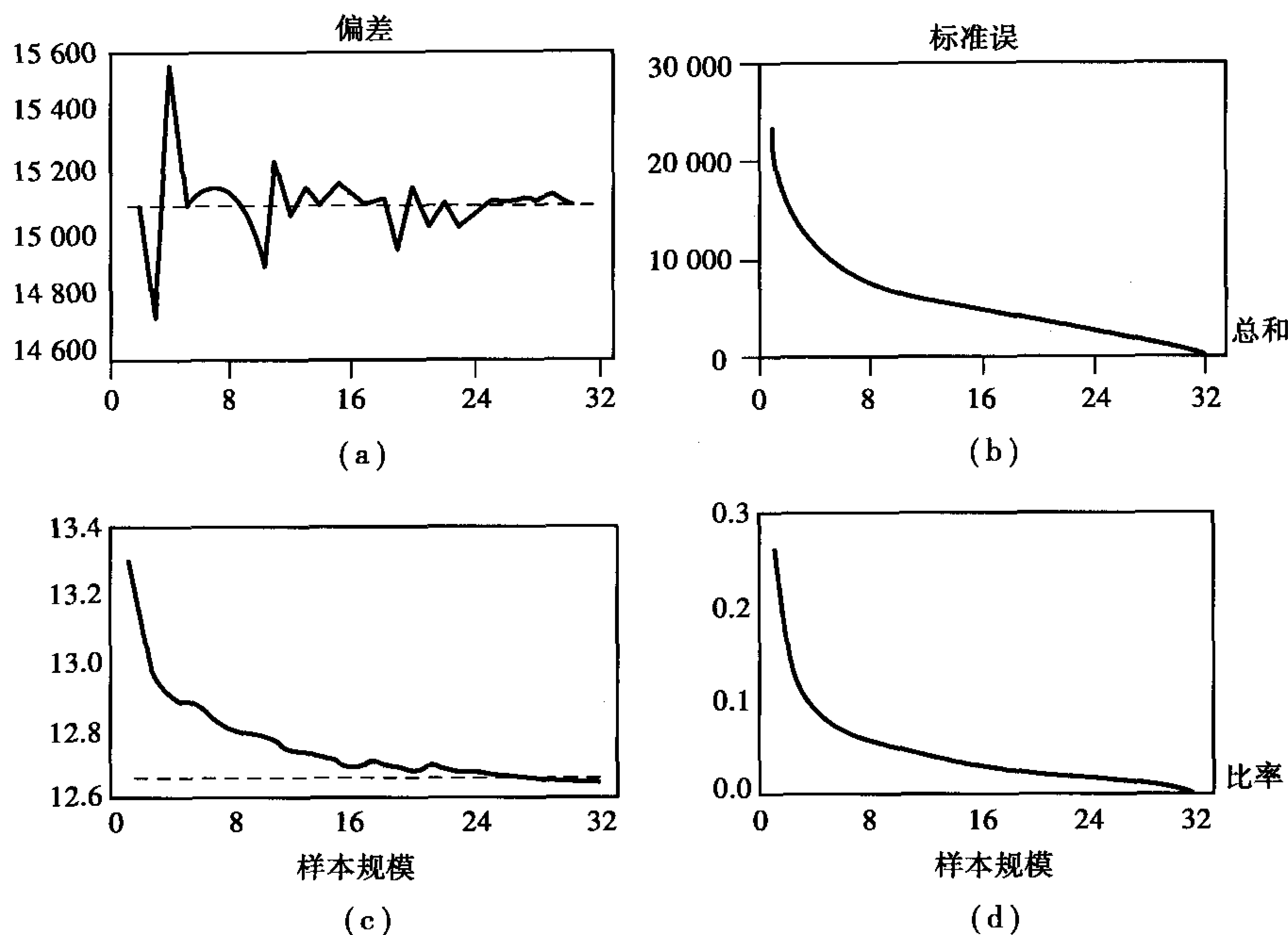


图 2.3　总和估计量 $\hat{t}$ 与比率估计量 $\hat{r}$ 的偏差、一致性与精度。
1991 年省级人口中各种样本规模的 SRSWOR 设计下 1 000 次模拟的蒙特卡洛均值 $\hat{t}_{mc}$ 与 $\hat{r}_{mc}$

DEFF 与抽样设计的效率

在前面,设计效应被定义为两个设计方差的比率。其中,分子是实际抽样设计的设计方差估计值,分母是样本元素相同的简单随机抽样的设计方差的估计值。这一定义最早由基什(Kish,1965,第 258 页)给出。他使用无放回式简单随机抽样为参照。更正式一些,定义总和估计值 $\hat{t}$ 在实际设计中的设计方差为 $V_{p(s)}(\hat{t})$,DEFF 参数如下:

$$\text{DEFF}(\hat{t}) = \frac{V_{p(s)}(\hat{t})}{V_{srs}(\hat{t})}。\tag{2.11}$$

在设计效应(2.11)中,估计值 $\hat{t}$ 被用于实际和参照设计中。在更加复杂的设计中(如章节 2.1 中),DEFF 使用更为普遍的公式,它能够引入基于设计的估计值 $\hat{t}^*$,而这一新引入的估计值与 SRSWOR 中的 $\hat{t}$ 是不同的。同时,在基什的定义中,SRSWOR 是参照。在实际中,这一定义的理解相对宽松。这是因为,当目标总体数目巨大而抽样比例 n/N 相对较小时,放回式或是无放回式简单随机抽样的结果是一样的。在大规模调查抽样中,经常就是这种情形。SRSWR 情形下的方差估计值在代数形式上比 SRSWOR 的要简约得多。从这种意义上,SRSWR 是一个更加方便的参照设计。这一点在调查分析的软件应

用中也得到了强调。

显然,如果实际抽样是 SRSWOR,则 DEFF = 1。对于放回式简单随机抽样(SRSWR),它的总体估计值 $\hat{t}$ 的设计方差是 $V_{srswr}(\hat{t}) = N^2(1-1/N)S^2/n$,其 DEFF 减少为:

$$\text{DEFF}(\hat{t}) = \frac{V_{srswr}(\hat{t})}{V_{srs}(\hat{t})} = \frac{N^2\left(1-\frac{1}{N}\right)\frac{S^2}{n}}{N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n}} = \frac{N-1}{N-n}。$$

当 $n \geqslant 2$ 时,这一 DEFF 总是大于 1,表示与 SRSWOR 相比,SRSWR 设计要相对低效。因此,SRSWR 的 DEFF 只取决于总体规模和样本规模。当总体数目巨大,而抽样比例 n/N 可以忽略时,则 DEFF 接近于 1。

在实际中,某一估计值的设计方差 $V_{p(s)}$ 与相应的 SRSWOR(或是 SRSWR)参照方差均是从已选样本中估算所得。因此,DEFF 必须从样本数据中得出。为了获取 deff,将使用方差的估计值。在下一个例子中,我们使用 1991 年省级人口数据中的资料来计算 DEFF 与 deff 数字。

范例 2.2

使用 SRSWOR 从 1991 年省级人口数据($N=32$)中抽取一个规模 $n=8$ 的样本。这一样本被假定使用了 SRSBE(贝努利抽样)与 SRSWR(放回式简单随机抽样)。我们计算失业人口总数 UE91 估计值 $\hat{t}$ 的 DEFF 参数。从总体中,我们知道标准差 $S=743$ 以及均值 $\bar{Y}=472$。因此,

$$\text{DEFF}_{srs}(\hat{t}) = 1(\text{由定义所得}),$$

$$\text{DEFF}_{srswr}(\hat{t}) = \frac{N-1}{N-n} = \frac{32-1}{32-8} = 1.29,$$

$$\text{DEFF}_{srsbe}(\hat{t}) = 1 - \frac{1}{N} + \frac{\bar{Y}^2}{S^2} = 1 - \frac{1}{32} + \frac{472^2}{743^2} = 1.37。$$

DEFF 参数显示,与参照 SRSWOR 设计相比,SRSWR 和 SRSBE 更加低效。对于 SRSBE 而言,增加的方差部分是因为随机样本规模原因。

我们从表 2.3 中显示的选出的样本中计算 deff 估计值。总体标准差的估计值为 $\hat{s}=1\ 355.615$,均值的估计值为 $\bar{y}=826.25$。如果把这一样本当成使用放回式简单随机抽样或是贝努利抽样得到,deff 的估计值是:

$$\text{deff}_{srs}(\hat{t}) = 1(\text{由定义所得}),$$

$$\text{deff}_{srswr}(\hat{t}) = \frac{N-1}{N-n} = \frac{32-1}{32-8} = 1.29,$$

$$\text{deff}_{srsbe}(\hat{t}) = 1 - \frac{1}{N} + \frac{\bar{y}^2}{\hat{s}^2} = 1 - \frac{1}{32} + \frac{826.25^2}{1\ 355.62^2} = 1.34$$

当然,SRSWR 中的 deff 估计值与参数 DEFF 相同。即使是 SRSBE 抽样,deff 估计值几乎与相应的 DEFF 参数相同。

我们讨论了无放回式简单随机抽样中总和、比率和中位值的设计方差以及方差估计值。对于总和的线性估计值 $\hat{t}$，我们推导了分析设计方差，得出了与相应的方差估算公式基本上相同的公式。对于作为非线性估计值的比率 $\hat{r}$，使用线性方法得出了近似设计方差，方差估计值与设计方差也非常相似。对于抗扰估计值中位值的设计方差，有替代近似估计值，但它们的实用性在小样本的情形下是各不相同的。

小　结

我们介绍了简单随机抽样，目的是为了增加在具体样本选择方案下，熟悉估算过程中最重要的概念。关键的统计概念有三个层次。第一层是诸如总和 T、比率 R 及中位值 M 的未知总体参数，它们是从抽取的样本中估算出来的。第二层是总体参数的估计值及其设计方差，包括设计参数和估计值抽样分布的其他特征。由于抽样的随机性，从总体中的复制样本中计算估计值的观测值产生了差异。而设计方差就是用来表示这些差异。它们也反映在估计值的抽样分布当中。看起来，把抽样分布的特征当成恰当的点估计、区间估计以及假设检验的基础是大有裨益的。抽样设计的效率由估计值的设计效应 DEFF 来表示。

实践中，只有实际抽取的样本用来估算。因此，第三层是总体参数的样本估计值与为求取标准误及置信区间的设计方差的估计值。使用样本中估算的设计方差和假定的简单随机抽样中相应的方差估计值，计算得出的 deff 估计值是相当重要的数值。

从 1991 年省级人口数据中，我们抽取了一个无放回式简单随机抽样的样本。根据这一样本，我们详细研究了涵盖这三个层次的总体、比率及中位值的估计值的特征。估计值 $\hat{t}$ 对应全省失业人口数 UE91 的总和 T；比率估计值 $\hat{r}$ 所对应的是全省失业率 R；中位值估计值 $\hat{m}$ 则对应每个自治市的平均失业人数 M。这些估计值涵盖了三种类型的估计值，线性、非线性及抗扰估计值。这种情形下，所有的 DEFF 数值与 deff 估计值相同。这是因为 SRSWOR 是设计效应计算中的参照。在其他抽样方案中，我们在稍后章节中会看到，效率怎样根据估计值和抽样设计而变化。在许多情形下，将得到不等于单位 1 的 deff 的估计值。

最后，需要注意，SRS 并不仅仅是一个用来演示调查抽样中抽样误差和其他关键概念的简单工具，也不仅仅是效率比较中的参照。简单随机抽样可以是复杂抽样调查中抽样设计的固有的部分，因而，它具有实际的价值。

2.4 系统抽样与组内相关

系统抽样是最常用的样本选取技术之一。将总体元素的排列或是计算机化的登记作为选取框架,从中系统地每隔 q 个总体元素选取一个样本元素。例如,许多人口登记是按姓氏的字母顺序排列的。第一个样本个案是在头 q 个元素里随机选取。剩下的则是每隔 q 个元素抽取,直至名单最后。我们花费大力气来讨论系统抽样的估算。这是因为它是一个很好的复杂性的例子。这样的复杂性,我们在某种设计情形下的估算过程中遇到,而这种情形涉及需要设计参数的设计方差的估计值。这里,设计参数是组内相关系数 ρ_{int}。更多的复杂性出现在设计方差的估算过程中。即使是对于简单的总和估计值,也没有已知的分析方差的估算公式。我们将推导几个方差估算公式。在它们中进行选择,得到的有关目标总体结构的信息是大有帮助的。

在某些情况下,系统抽样比简单随机抽样更为有效。例如,当总体框架的排列顺序与研究变量的取值之间有着某种关系时,这种情况就会发生。最常见的情形是,总体已经分层,或是总体的顺序有一定的趋势,或是有一个时间序列。所有这些情形都可以通过适当的调整达到。在某些情形下,周期性可能非常有害,特别是当协变与抽样间隔相同时。对有关总体结构的预先的了解,对于获取有效的估算是大有帮助的。

抽取样本

假定,需要从有 N 个元素的固定总体中抽取规模为 n 的系统样本。有几种抽取的方法。最常见的方法是抽取一个抽样间隔为 $q = N/n$,规模为 n 的样本。其他的,是抽取两个或是更普遍的 m 个重复的系统样本,每个样本有n/m 元素,而抽样间隔为 $m \times q$。这种方法适用于方差估算需要使用所谓的重复技术时。

让我们考虑单一随机起点的系统抽样。第一个任务是,将总体框架中元素标注连续的数字 $1,2,\cdots,q,q+1,\cdots,N-1,N$。其中,$q = N/n$ 是抽样间隔。当 q 不是整数时,除去一个以外的所有的抽样区间可以有相同的长度。选择的程序如下。在 1 到 q 之间以 $1/q$ 的概率随机选出一个整数,让它为 q_0,样本中的元素的数字标签为 $q_0,q_0+q,q_0+2q,\cdots,q_0+(n-1)q$。每个抽样区间有一个元素被选中。

单一随机起点的另一种选择方法是,在区间[1,N]中随机选取一个整数。让它为 Q_0。从它开始,以 q 为抽样间隔,向前和向后选取。系统样本的组成是:$\cdots,\ Q_0-2q,\ Q_0-q,\ Q_0,\ Q_0+q,\ Q_0+2q,\ \cdots$。系统抽样还有一种方法,是

将框架中的元素首尾相连成为一个封闭的连环。从随机起点 Q_0 开始,依次抽取得到 Q_0+q, Q_0+2q,…,直到框架末尾。然后,又从框架的起点继续抽取。连环将在抽取到 n 个元素后结束。这些方法都可以选出规模为 n 的样本,它们在估算过程中是等价的。

在重复系统抽样中,使用的是多元随机起点。计划中的样本规模 n 被分配到 m 个次级样本中,每一个有着相同的样本规模 n/q 的次级样本中的抽样间隔为 $m\times q$。在每一个次级样本的第一个抽样区间中,使用无放回式方法选择一个整数作为随机起点。选取这一起点的方法是上述第一种方法。这一过程得到总数为 n 个不同元素的一系列同等规模的重复系统样本。

系统抽样中,不同样本的数目较小。当样本间隔为 $q=N/n$ 时,共有 q 个不同的系统样本。所以,选出样本 s 的概率为 $p(s)=1/q$。当每一个抽样区间中一个元素被抽中时,第 k 个总体元素的选中概率为 $\pi_k=1/q=n/N$。它也等于无放回式简单随机抽样中的选中概率。所以说,系统抽样也是无放回式等概抽样设计中的一种。

估　算

系统抽样选取样本的简便在估算过程中消失了。使用简单随机抽样中相应的估算公式,可以简便地计算出总和 T、比率 R 及中位值 M 的点估计。但是,从所选样本中无法分析地估算设计方差,这需要近似算法。这是因为,每个抽样区间内,只有一个总体元素被选中,因而样本中没有抽样间隔内离异的信息,而这些信息是分析估计方差所需的。这个问题可以用下面总和 T 的估算公式来显示:

$$\hat{t}=N\sum_{k=1}^{n}\frac{y_k}{n},\tag{2.12}$$

这与 SRSWOR 等式(2.1)相同。在系统抽样情形下,$\hat{t}$ 的设计方差如下,

$$V_{sys}(\hat{t})=N^2\sum_{j=1}^{q}\frac{(\bar{Y}_j-\bar{Y})^2}{q},\tag{2.13}$$

其中,$\bar{Y}_j$ 是第 j 个系统样本的均值,$\bar{Y}$是总体均值。离差的大小取决于在多大程度上 q 个样本均值$\bar{Y}_j$ 围绕总体均值$\bar{Y}$的变动。如果每一个样本都与总体的成分相近,则设计方差较小,而总和的估算较为有效。但是,如果各个样本均值变动较大,则会得到较大的设计方差。这种情形可以看成是在系统样本间和系统样本内分解总的离差。我们将在组内相关中作进一步讨论。

在实际中,只抽取了一个系统样本,而使用替代——多少含有偏差的——方差估计值 $\hat{v}_{sys}(\hat{t})$来近似估算设计方差。近似方差估算公式的确定,要么取决于有关总体框架的辅助信息,要么运用某种诸如样本重复使用或是使用重复系统样本的解决方法。式 2.14 到式 2.18 中介绍 5 个近似的方差估算公式。

1. 随机排列的总体。很自然,我们经常假定总体中研究变量的取值是随

机排列的。如果这一模型是正确的，无放回式简单随机抽样下的方差估计公式如下：

$$\hat{v}_{1.sys}(\hat{t}) \approx \hat{v}_{srs}(\hat{t}) = \frac{N^2\left(1 - \frac{n}{N}\right)\hat{s}^2}{n}, \quad (2.14)$$

它在实际的系统样本下是无偏的。这一模型虽然很少正确，但在诸如以姓氏字母排序的人口登记中看起来是符合实际情况的。

2. 隐性分层总体。总体中的元素根据某一变量的取值调整过。例如，在人口登记中，人们根据性别先女后男来排列。这种分层叫做隐性分层。对应的近似方差估算公式需要用到相邻取值间的差 $a_i = y_i - y_{i-1}$，其公式如下：

$$\hat{v}_{2.sys}(\hat{t}) \approx N^2\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)\sum_{i=2}^{n} \frac{a_i^2}{2(n-1)}。 \quad (2.15)$$

另外，也可以直接使用按比例分配的分层随机抽样的方差估算公式，这一公式要使用在各个隐性分层中 SRSWOR 条件下的等式2.6。我们将在章节 3.1 中介绍 $\hat{v}_{2.sys}(\hat{t})$。

3. 自相关总体。这种情形出现在超总体的机制中。总体中相隔 q 单位的元素每两两产生相关系数 ρ_q。它与时间序列分析中的自相关类似。它的取值假定为正；若不是，则需要使用其他近似方法。可以从抽取的样本中估算出自相关系数，并将它用作方差公式 $\hat{v}_{sys}$ 中的一个校正因子：

$$\hat{v}_{3.sys}(\hat{t}) \approx \frac{N^2\left(1 - \frac{n}{N}\right)}{n}\left(\frac{\hat{s}^2}{n}\right)\left[1 + \frac{2}{\log(\hat{\rho}_q)} + \frac{2}{\hat{\rho}_q^{-1} - 1}\right], \quad (2.16)$$

其中，$0 < \hat{\rho}_q < 1$ 是自相关的估计值。当自相关大于 0 时，中括号里的部分小于 1；随着 $\hat{\rho}_q$ 的增加，该部分进一步减小，并趋近于 0。因此，高自相关系数增加效率。

4. 样本重复使用。把原样本分成两个或多个等规模的互不交叉的系统样本。通过这 m 个次级样本之间的差异来估算设计方差如下：

$$\hat{v}_{4.sys}(\hat{t}) \approx N^2\left(1 - \frac{n}{N}\right)\sum_{l=1}^{m} \frac{(\bar{y}_l - \bar{\bar{y}})^2}{m(m-1)}, \quad (2.17)$$

其中，$\bar{\bar{y}} = \sum_{l=1}^{m} \bar{y}_l/m$ 是 m 个次级样本均值的均值。在式2.17 中，$\bar{y}$ 可以用来替代 $\bar{\bar{y}}$。其他重复使用样本的技术，如脱靴、折刀以及平衡半样本等，是方差估算中可能的方法。重复使用样本的方法将在第 5 章中得到进一步讨论。

5. 重复系统样本。这种方法与上述将原样本分成两个或是多个次级样本的方法类似。但是，这里是在样本抽取之前。使用无放回式方法抽取两个或是多个重复的系统样本。这 m 个样本间的差异提供了估算设

计方差的机会。近似方差的公式与前面方法的公式相同，即，

$$\hat{v}_{5.sys}(\hat{t}) = \hat{v}_{4.sys}(\hat{t})。 \tag{2.18}$$

这5种方差估计的方法都是近似的。因此，它们的统计特征取决于各自模型的假设是否有根据或是对原样本的分割是否成功。当然，在实际中不可能确定这些的正确性。但是，我们可以通过1991年省级人口数据来评估这些方差估算的有效性。因为，我们可以计算设计方差 V_{sys} 以及相应的作为设计参数的组内相关系数 ρ_{int}。

范例 2.3

我们使用1991年省级人口数据的系统样本来近似地计算方差。有两种方法从总数为32个自治市的1991年省级人口数据中抽取规模为8（$n=8$）的系统样本：

1. 全省被分成8个抽样区间，每个区间有4个自治市。从中选出一个样本。比如，选取每一个抽样区间的第一个自治市。这样，样本规模为8个自治市。
2. 全省被分成4个抽样区间，每个区间有8个自治市。使用无放回式方法，从中选出两个平行的系统样本。比如，选取每一个抽样区间的第一个和第五个自治市。这样，样本中就有了两个互不交叉的，规模为4个自治市的复制系统样本，而样本总规模为8个自治市。

在抽取实际样本中，假定使用了这两种方法。表2.5给出了抽取的数据。在表2.1中隐性分层是根据城市登记的自治市的顺序——人口密集的城镇自治市在先，农村自治市在后。这种登记框架下的系统抽样，按照城镇与农村各自的比例在各个层级中抽取自治市。这一抽样方法与按比例分配的分层抽样的结果是相同的。分层抽样将在章节3.1中讨论。

表2.5　1991年省级人口中的一个系统样本(样本设计标识表示隐性分层)

样本设计标识			元素标签	研究变量	
STR	CLU	WGHT		UE91	LAB91
1	1	4	Jyväskylä	4 123	33 786
1	5	4	Saarijärvi	721	4 930
2	9	4	Joutsa	194	2 069
2	13	4	Kinnula	129	927
2	17	4	Korpilahti	239	2 144
2	21	4	Leivonmäki	61	573
2	25	4	Petäjävesi	262	1 737
2	29	4	Säynätsalo	166	1 615

抽样比例：层级1=0.2，层级2=0.25。

根据这一抽取的数据，我们计算了全部5种方差近似估计值。为了计算

分层假设中的方差估计值，当自治市为城镇时，层级标签取值 STR = 1；自治市为农村时，取值为 STR = 2。同样的，和简单随机抽样中一样，整群标签（CLU）取值为元素标注号码。与简单随机抽样的情形相同，按比例抽样中的元素的权重是一常数，在这里权重 WEIGHT = 4。每个层级均给出了抽样比率，但它们同为 0.25。

表 2.6 给出了隐性分层情形下的估算结果与相应的参数取值。$\hat{t}$，$\hat{r}$ 与 $\hat{m}$ 的点估计值与在 SRSWOR 设计中的所得相等，但方差估计值有差异。这里使用了 $\hat{v}_{2.str}(\hat{t})$。总和与中位值的 deff 估计值比 1 小得多。与无放回式简单随机抽样中计算得来的方差估计值相比，使用隐性分层的系统抽样中的方差近似估算要有效得多。比率的 deff 估计值大于 1，显示使用隐性分层没有任何提高。

表 2.6 使用隐性分层从 1991 年省级人口抽取的系统样本的估计值

统计量	变 量	参 数	估计值	s. e	c. v	deff
总和	UE91	15 098	23 580	11 802	0.50	0.76
比率	UE91, LAB91	12.65%	12.34%	0.33%	0.03	1.29
中位值	UE91	229	198	27	0.14	0.21

让我们更进一步来讨论总和 $\hat{t}$ 的方差近似估算。当然，总和 T 的点估计值在所有近似估算中相同，为 $\hat{t}$ = 23 580。在分层的假设下，有两个方差估计：一是基于隐性分层的 $\hat{v}_{2.str}$，另一是基于相邻取值间的差的 $\hat{v}_{2.sys}$。放到一起，有下面近似方差的估计：

$$\hat{v}_{1.sys}(\hat{t}) \approx N^2\left(1-\frac{n}{N}\right)\frac{\hat{s}^2}{n} = 13\ 549^2 \qquad \text{deff} = 1.00$$

$$\hat{v}_{2.sys}(\hat{t}) \approx N^2\left(1-\frac{n}{N}\right)\left(\frac{1}{n}\right)\sum_{i=2}^{n}\frac{a_i^2}{2(n-1)} = 13\ 220^2 \qquad \text{deff} = 0.95$$

$$\hat{v}_{2.str}(\hat{t}) \approx \sum_{h=1}^{2}\hat{v}(\hat{t}_h) = 11\ 802^2 \qquad \text{deff} = 0.76$$

$$\hat{v}_{3.sys}(\hat{t}) \approx N^2\left(1-\frac{n}{N}\right)\left(\frac{\hat{s}^2}{n}\right)\left[1+\frac{2}{\log(\hat{\rho}_q)}+\frac{2}{\hat{\rho}_q^{-1}-1}\right] = 8\ 224^2$$

$$\text{deff} = 0.35$$

$$\hat{v}_{4.sys}(\hat{t}) = \hat{v}_{5.sys}(\hat{t}) \approx N^2\left(1-\frac{n}{N}\right)\sum_{l=1}^{m}\frac{(\bar{y}_l-\bar{\bar{y}})^2}{m(m-1)} = 12\ 959^2$$

$$\text{deff} = 0.87$$

在近似方差估计值中，基于 SRSWOR 假设的 $\hat{v}_{1.sys}$ 的数值最大。其他的多少在其之下。这种情形显示了，系统抽样比简单随机抽样更为有效。最为有效的近似方法是自相关模型，其得出 deff = 0.35。这一模型的假设是一个自相关的超总体，而固定总体是它的一个实际例子。设计效应为 DEFF = 0.55，也证实了这一点。

为了评价方差估算的结果，我们可以研究组内相关系数ρ_{int}——系统抽样中的单一设计参数——的特征以及抽样方案的效率。同样的，它也显示登记框架中的顺序调整与组内相关系数之间是何种关系。

组内相关

系统抽样是第一个含有设计参数的设计例子。这个叫做组内相关系数ρ_{int}的参数，被包括在估算公式的设计方差V_{sys}中。组内相关的大小及其对于方差估计值的效应，部分取决于抽样间隔，部分取决于总体框架中研究变量的取值是否含有连续的顺序。在系统抽样中，$\hat{t}$的设计方差在式2.13中为$V_{sys}(\hat{t})=N^2\sum_{j=1}^{q}(\bar{Y}_j-\bar{Y})^2/q$，其设计方差可以写成，

$$V_{sys}(\hat{t}) = \sum_{j=1}^{q}\frac{(N\bar{Y}_j - N\bar{Y})^2}{\frac{N}{n}} = N\times\sum_{j=1}^{q}n\times(\bar{Y}_j-\bar{Y})^2。\qquad(2.19)$$

让我们进一步分析设计方差（式2.19）。首先，我们将总体方差分解成系统样本间与系统样本内的离异。这正如标准的一元方差分析。使用ANOVA的术语，我们有：

$$SST = SSW + SSB,\qquad(2.20)$$

其中，SST是离差总和，SSW是组内离差，而SSB是组间离差。式2.20中的分解可以写成：

$$\sum_{k=1}^{N}(Y_k-\bar{Y})^2 = \sum_{j=1}^{q}\sum_{k=1}^{n}(Y_{jk}-\bar{Y}_j)^2 + \sum_{j=1}^{q}n(\bar{Y}_j-\bar{Y})^2。\qquad(2.21)$$

因此，设计方差另外的形式是$V_{sys}(\hat{t})=N\times SSB$。

使用式2.20中离差总和的分解，组内相关系数被定义为：

$$\rho_{int} = 1-\frac{n}{n-1}\times\frac{SSW}{SST}。\qquad(2.22)$$

当均值间的离差为0，或是$SSB=0$，组内相关达到其最低值，$-1/(n-1)$；相应的，当$SSW=0$时，它达到最大值，$\rho_{int}=1$。

更进一步，我们可以将总和估计值的方差写成以下的形式：

$$V_{sys}(\hat{t}) = N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n}[1+(n-1)\rho_{int}],\qquad(2.23)$$

或是写成SRSWOR的设计方差与一个校正因子的乘积。这一校正因子含有组内相关系数。

$$V_{sys}(\hat{t}) = V_{srs}(\hat{t})\times[1+(n-1)\rho_{int}]。$$

因此，设计效应是，

$$\mathrm{DEFF}_{sys}(\hat{t}) = \frac{V_{sys}(\hat{t})}{V_{srs}(\hat{t})}\approx 1+(n-1)\rho_{int}。\qquad(2.24)$$

与简单随机抽样相比，系统抽样

1. 更有效，当 $-1/(n-1)<\rho_{int}<0$ 时，
2. 同等有效，当 $\rho_{int}=0$，或是
3. 相对低效，当 $0<\rho_{int}<1$ 时。

这可以被理解为，抽样区间异质性越高（即组内相关为负），系统抽样的效率越高。所以，系统抽样中，设计参数 ρ_{int} 与总体框架的调整顺序有联系。这样的联系可以在实际中得以有效地利用。

范例 2.4

1991 年省级人口数据中组内相关系数的计算。现在，我们将在系统抽样的情形下估算 UE91 总和，并计算 1991 年省级人口数据中组内相关系数。我们使用单一的含有 8 个元素的系统样本来计算系统抽样。表 2.7 给出了离差总和的分解（式 2.21）。

表 2.7　总体 ANOVA 表：系统抽样 $q=4, n=8$

离异来源	df	平方和	MSE
样本间	3	$SSB=9.18\times10^5$	$MSB=3.06\times10^5$
样本内	28	$SSW=162.14\times10^5$	$MSW=5.79\times10^5$
合计	31	$SST=171.32\times10^5$	$S^2=5.53\times10^5=743^2$

因此，组内相关系数为：

$$\rho_{int}=1-\frac{n}{n-1}\times\frac{SSW}{SST}=1-\frac{8}{8-1}\times\frac{162.14\times10^5}{171.32\times10^5}=-0.082。$$

因为组内相关系数为负，与无放回式简单随机抽样相比，系统抽样更为有效。因此，设计效应为：

$$DEFF_{sys}(\hat{t})\approx1+(n-1)\rho_{int}=1+(8-1)\times(-0.082)=0.426,$$

显示系统抽样在这里相当有效。

接下来，我们进一步考察系统抽样在不同模型假设下或是不同的总体的调整顺序的效率。早先在一个给定的样本中也讨论过这一假设。我们现在使用相应的设计方差。

范例 2.5

隐性分层与 DEFF。在 1991 年省级人口数据中，城镇自治市在先，然后是农村自治市。它们都是按字母顺序排列。因此，名单的顺序牵涉到两个隐性层级。在第一个层级中，城镇自治市有着较多的人口以及失业人数。所以，随着自治市标签的顺序，有一个轻微的下降趋势。相应的散点图（图 2.4）显示研究变量 UE91 取决于总体中元素的调整顺序间的关系。

UE91 的取值有赖于名单顺序的事实，对于选择合适的方差估计值有着意义。

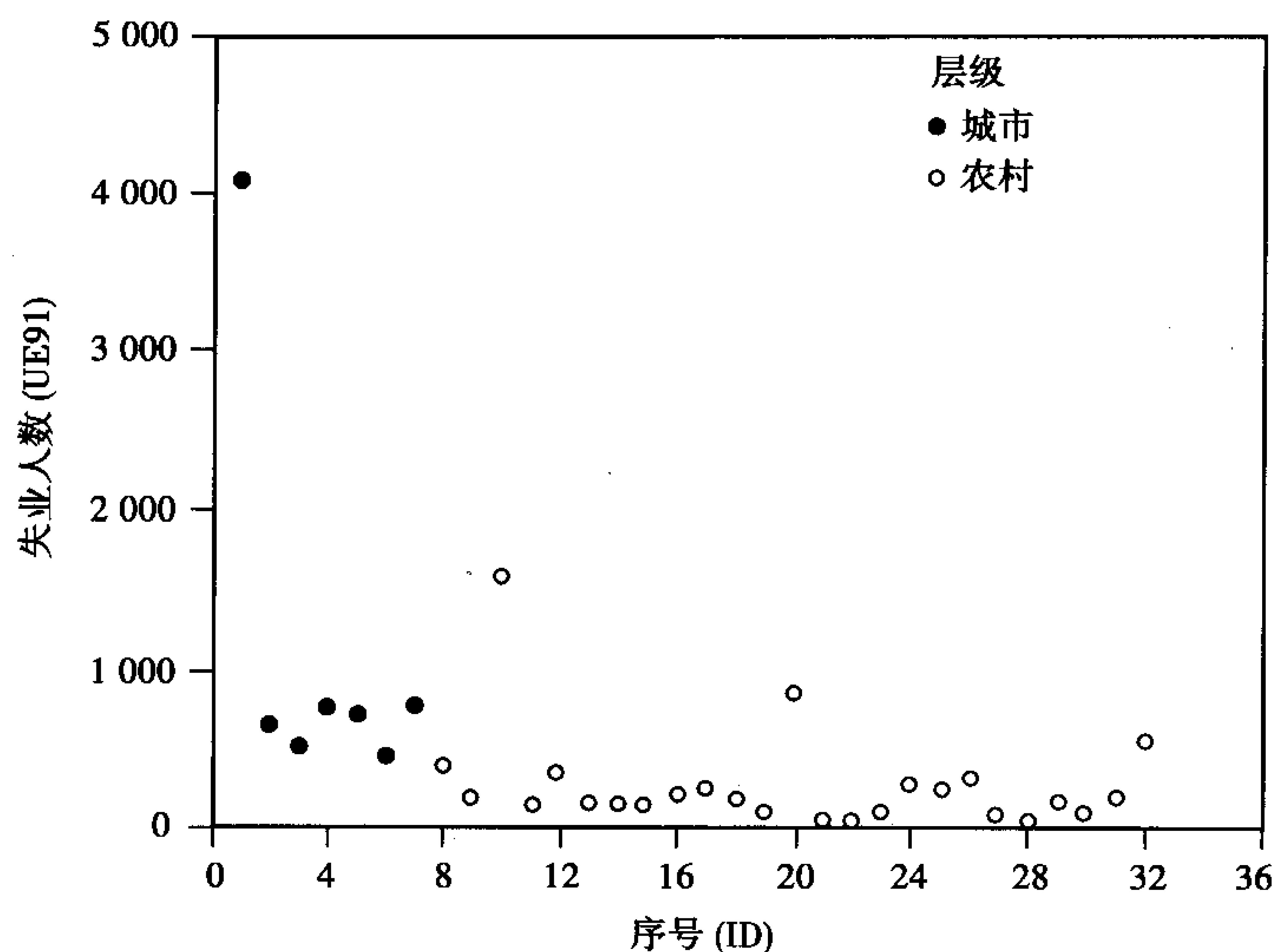


图 2.4　1991 年省级人口 UE91 对序号(ID)图(标明了两个隐性分层的两个层级)

1. 分布图明白地显示相邻的顺序并非随机。因此,这一样本不应该被看成是一个简单随机样本。在前面,我们知道 $\text{DEFF}_{sys}(\hat{t}) = 0.554 < 1$。所以,SRSWOR 的设计方差 $V_{srs}(=7\ 283^2)$ 可能夸大了设计方差 $V_{sys}(=5\ 420^2)$。

2. 在登记表中,总体按照层级的顺序依次排列。对于隐性层级而言,可以计算出层级规模以及 UE91 的均值,如下表所示:

层级	序号	规模	均值
1. 城市	1 ~ 7	7	1 146
2. 农村	8 ~ 32	25	283
全体	1 ~ 32	32	472

系统抽样揭示了这些隐性层级,并抽取了相应的按比例分层的样本(STR)。如果已知层级权重,这一样本可以被当成一个后续分层样本来分析。章节 3.3 将讨论这些。

分层抽样情形下的设计效应是:

$$\text{DEFF}_{sys,str}(\hat{t}) = \frac{6\ 251^2}{7\ 283^2} = 0.737,$$

所以,分层使得估算更加有效。因此,这一近似估算也夸大了真实的设计方差。

3. 研究变量与顺序数字之间有简单的线性关系。这一关系可以用简单的线性回归模型来表示:

$$Y_k = 1\ 070.72 - 36.30 \times ID_k。$$

这个模型的多元相关系数的平方是 $R^2=0.21$。在实际估算中将这一回归模型当成辅助信息,我们可以使用回归估算(参见章节 3.3)。比如,回归估算中的设计效应是:

$$\mathrm{DEFF}_{srs,reg}(\bar{y}) \approx 1-R^2=1-0.21=0.79,$$

落入区间 $0.554<0.79<1$,而 0.554 是系统抽样情形下 $\hat{t}$ 的实际的 DEFF。

4. 自治市的排列顺序里还包含了相邻自治市间的自相关。使用抽样间隔 $q=4$ 作为滞后,自相关系数为 $\rho_4=0.090\ 85$。在这一自相关系数下得到的设计效应是:

$$\mathrm{DEFF}_{srs,autocor}(\bar{y}) \approx \frac{4\ 405^2}{7\ 283^2}=0.366,$$

与系统抽样情形下实际的设计效应 0.426 非常接近。自相关唯一的短处是,总体框架含有对应于抽样间隔的协变。这里的情形不是这样的。

5. 登记表的预先调整与系统抽样的效率。登记的框架通常是以计算机化的数据形式出现的,因而可以根据合适的变量加以调整。调整的过程影响抽样区间的内容,但并不是如期望地损害估算的效率。比如,使用失业人数把 1991 年省级人口数据调整成单一下降趋势的排列顺序。另外,我们也调整抽样区间的顺序,使得失业人数在每第二个区间内呈下降趋势,但在另一个区间内呈上升趋势。这样,我们得到系统抽样情形下总体框架的最佳顺序。相应的设计方差为 $V_{sys,opt}(\hat{t})=2\ 348^2$,而 DEFF $=0.104$,显示顺序调整有着显著的优势。所以,调整顺序以某种隐性分层常见于大规模调查中。

小　结

从一个计算机化的登记框架中,很容易使用系统抽样。因而在实际中,它运用广泛。但是,难点是在系统抽样的情形下设计方差的估计值的估算过程。一种解决的方法是,使用总体框架中已有的辅助信息。如果登记表格里元素的排列完全是随机的假设是合理的,简单随机抽样的估算公式是可以使用的。但是,如果诸如隐性分层、趋势或时效性的某种研究变量的结构存在于登记表当中,在估算近似方差的过程中使用这些信息则更为有效。在我们的例子中,与使用 SRSWOR 得到的估计值相比,使用这些近似估算公式得到的估计值与实际的设计方差更为接近。这是因为总体中存在某种结构。特别是在大型系统抽样中,很值得尝试重复使用样本的技术,它可以导出其他近似的方差估算公式。沃尔特(Wolter,1985)给出了一个更有体系的系统抽样中方差估算的研究。他指出,尝试其他方差估算方法是值得的,这样可以找出最适合当前情况的方法。

因为系统抽样在实际中的流行以及它涉及一个有趣的设计参数——即,

组内相关系数，我们对它进行了广泛的讨论。这一设计参数并非十分关键。但是，它对于方差估算、抽样误差的标明、置信上限下限以及检验的规模，有着特殊的效应。因此，我们通过一个模型辅助估算过程给出和补充了近似方差估算的主线。

2.5 概率对应规模抽样

有时会遇到这样一种情况，总体中含有一些元素，它们在研究变量上的取值巨大。这在商业调查中经常出现。特别对于估算总和而言，选中概率取决于总体元素规模的抽样技术是非常适合这种情形的。如果规模变量与研究变量高度相关，方差极有可能得到降低。由于这种技术是基于选中概率与总体元素的相对规模成比例，它被称为概率对应规模抽样(PPS)。

在 PPS 抽样中，选中概率根据元素的相对规模而不同。总体元素的规模由辅助变量 z 来表示。我们假定已知总体元素 k 的辅助变量取值 Z_k。相对规模等于商值$p_k = Z_k/T_z$，其中 T_z 是总体中辅助变量的总和，或者更精确些，$T_z = \sum_{k=1}^{N} Z_k$。通常使用的规模的变量是那些能够自然测量总体中元素规模的变量。比如，在商业调查中，公司的员工人数是一个方便的度量规模的变量。又如，学校调查中学生总数也是一个较好的度量规模的变量。

选取辅助变量 z 的原因是它与研究变量 y 的变动相似。更准确些，规模变量 z 与研究变量间的比率尽可能接近一常数。这是因为，PPS 中的效率取决于比率 Y_k/Z_k 在多大程度上对于所有的总体元素保持为常数 C。当这一比率接近于常数时，估计值的设计方差就较小。

在 PPS 抽样中，选中概率 π_k 与元素的相对规模 $p_k = Z_k/T_z$ 成比例。而抽取的元素的个体权重取决于这些相对规模的倒数。可以使用放回式或是无放回式方法抽取 PPS 样本。在放回式样本中，选中概率的计算相对简单些。在无放回式 PPS 抽样中，获取这些概率非常复杂。这是因为抽取第一个元素后，剩下的$(N-1)$个元素的相对规模起了变化，因而需要计算新的选中概率。人们发展了多种技术来克服这一困难。当规模变量合适时，特别是对于估算总和的例子，PPS 抽样相当有效。

抽取样本

我们提出几个使用概率对应规模的抽样方案。起点是了解每个总体元素的辅助变量 z 的取值，以便我们计算选中概率。选中概率 π_k 与元素 k 的相对规模 Z_k/T_z 成比例。比如，在简单随机抽样的简便例子中，元素 k 的相对规模是 $p_k = 1/N$。数值 $1/N$ 被称为元素 k 的单一抽取选中概率。样本规模为 n 中

元素的选中概率为 $\pi_k = n \times p_k = n/N$。但在 PPS 抽样中,选中概率 π_k 并不恒定。因此,与简单随机抽样和系统抽样相反,它并不是一个等概抽样设计。

在实际中,PPS 样本的抽取可以根据总体元素的相对规模,或是根据规模变量的累积和。第 k 个元素的累计和是:

$$G_k = \sum_{j=1}^{k} Z_j,\ k = 1,\ \cdots,\ N,\ G_N = T_z。$$

自然数 $[1, G_1]$ 与第一个总体元素相联,数值 $[G_1 + 1,\ G_2]$ 与第二个元素相联。更普遍的,第 k 个元素所对应的数值在区间 $[G_{k-1} + 1,\ G_k]$ 之中。样本抽取的过程有赖于这些数字。

我们讨论 5 种具体的 PPS 抽样方案。它们是与贝努利抽样类似的泊松抽样、放回式或是无放回式累积总和法、非等概系统抽样,以及拉奥-哈特利-科克伦方法(RHC method;Rao et al. ,1962)。这些方法中,将详细讨论放回式累积总和法与非等概系统抽样。在例子中,变量 HOU85 度量总体元素的规模。它是基于登记的,表示总体自治市中家庭数目。

泊松抽样　这一抽样方案使用依名单顺序抽取程序。首先,计算选中概率 $\pi_k = n \times Z_k/T_z$。接下来,从统一分布$(0,\ 1)$中抽取独立的随机数字 $\varepsilon_1, \cdots, \varepsilon_k, \cdots, \varepsilon_N$。当 $\varepsilon_k < \pi_k$ 时,元素 k 就被选取了。对于总体中的每个元素 $k = 1, \cdots, N$ 都依次使用这一过程。

显然,在泊松抽样中,样本规模并不是预先确定的,而是一个随机变量。样本规模的期望值为 $E(n_s) = \sum_{k=1}^{N} \pi_k$。泊松抽样有时在商业调查中作为配位样本(Ohlsson,1998)。

放回式 PPS 抽样(PPSWR)　由于其相应的设计方差公式容易处理,放回式抽样在评估估计值的统计特征时有自己的优势。放回式 PPS 抽样与放回式简单随机抽样很近似。两种方法的不同之处在于抽取数字与总体元素对应的方式。在简单随机抽样中,单一的自然数 $1, \cdots, k, \cdots, N$ 与总体元素相对应。相对的,在 *PPS* 抽样中,从数字 $1, \cdots, G_k, \cdots, G_N$ 中的区间与总体元素相对应。其中,G_k 是累计和。

放回式 PPS 抽样首先从区间 $[1, G_N]$ 中选出一个随机数。将这一数字与总体元素所对应的数字相比较。如果一个元素的选择区间包含这个随机数,那么,它就被选中了。一个元素的单一抽取选中概率为 $p_k = Z_k/T_z$。在所有抽样中,样本中元素 k 的选中概率为 $\pi_k = n \times p_k$。应当注意到,放回式抽样中,同一总体元素可能被抽中几次。对于规模巨大的总体元素而言,更有可能是这样的,因为它们的选中概率也较大。

无放回式 PPS 抽样(PPSWOR)　使用无放回式抽样时,产生了一个计算选中概率的问题。选取第一个元素时,单一抽取概率正好是 $p_k = Z_k/T_z$。抽取第一个元素之后,单一抽取概率变化了。这是因为,剩下的$(N-1)$个元素的

总和 T_z 减少了。特别是对于大规模样本,选中概率的计算相当复杂。因为这一原因,人们发展了大量的替代无放回式抽样技术来克服这一困难。比如,总体可以分成若干互不交叉的次级总体或是层级。然后,与布鲁尔(Brewer,1963)与默西(Murthy,1957)的方法相同,使用无放回式方法在各个层级中抽取两个元素。也可以从各个次级中抽取多于两个的元素,正如萨姆福德(Sampfard,1967)的方法。我们将详细讨论使用无放回式 PPS 抽取两个或更多元素的两种方法。

系统 PPS 抽样(PPSSYS) 在无放回式概率对应规模抽样中,这种方法简便易行。在这种方法中,系统抽样和概率对应规模抽样的特征组合成一体。在普通系统抽样中,抽样间隔取决于商值 $q = N/n$。在系统 PPS 抽样中,抽样间隔等于 $q = T_z/n$。与普通单一随机起点系统抽样相同,我们首先从封闭区间 $[1, q]$ 中选出一个随机数,让它为 q_0,则样本中 n 个选中数字为:

$$q_0, q_0 + q, q_0 + 2q, q_0 + 3q, \cdots, q_0 + (n-1)q。$$

在每一次选取中,当名单中的第一个总体元素的累积规模 G_k 大于或是等于其选中数字时,该元素即被选入样本。使用这种方法,样本中第 k 个元素的选中概率为 $\pi_k = n \times p_k$。

拉奥-哈特利-科克伦 PPS 方法(RHC method) 首先,使用规模变量 z,将总体分成 n 个次级总体 $N_1, N_2, \cdots, N_g, \cdots, N_n$,使得次级总体 g 中的规模变量总和 T_g 接近于 T_z/n。次级总体中元素的数目可以不同。接下来,从每一个次级总体中,按概率对应规模抽取一个元素。因而,元素 k 的选中概率为 $p_k = Z_k/T_g$。RHC 方法操作简便,适合于多种 PPS 抽样的情形。

估 算

放回式与无放回式情形下的估算应当分别考量。在放回式抽样中,元素的单一抽取选中概率保持为一个常数(即,等于该元素的相对规模 p_k)。但在无放回式抽样中,剩余元素的选中概率在每次抽取后产生变化。特别是对于方差估算,这造成了难题。为了介绍 PPS 抽样下估算的基本原则,我们将只讨论放回式的情形。作为近似的例子,将在例子中广泛使用的 PPSSYS 方法,也被简化为放回式的方法。

为了建立一个估算公式,需要总体元素 k 的相对规模 p_k。利用规模变量 Z_k,相对规模为:

$$p_k = \frac{Z_k}{\sum_{k=1}^{N} Z_k} = \frac{Z_k}{T_z}。$$

数值 p_k 也是元素 k 的单一抽取选中概率。随后,规模为 n 的样本中,元素 k 的选中概率 π_k 可以写成:

$$\pi_k = n \times p_k = n \times \frac{Z_k}{T_z}。$$

选中概率应当满足条件 $\pi_k \leq 1$。对于 $n = 1$ 的例子，总体中的每一个元素均满足这一要求。当 $n > 1$ 时，并且某些总体取值 Z_k 过大时，某些元素的选中概率会大于1，$n \times Z_k / \sum_{k=1}^{N} Z_k > 1$。这样的冲突在实际中可能遇到，但幸运的是可以克服。一种可能是，当 $nZ_k > \sum_{k=1}^{N} Z_k$ 时，将元素 k 的选中概率设定为 $\pi_k = 1$，即确定要抽中元素 k。实际中，这些元素组成单个元素的层级。而剩余的元素，π_k 与规模变量成比例。比如，总体中只有一个元素的 k' 的数值过大，那么设定 $\pi_{k'} = 1$，其余 $N-1$ 个元素的选中概率为：

$$\pi_k = (n-1)\frac{Z_k}{\sum_{k=1}^{N} Z_k - Z_{k'}}, k \neq k',$$

这就保证了 $\pi_k \leq 1$。范例2.8 中有这样的实际应用。

两个著名的 PPS 样本总和估计值——霍维茨-汤普森或是 HT 估计值与汉森-赫维茨或 HH 估计值就是根据这些概率数值得出。让我们推导出总和 T 的这些估算公式。在无放回式 PPS 抽样中，T 的 HT 估计值为（Horvitz and Thompson 1952）：

$$\hat{t}_{HT} = \sum_{k=1}^{n} \frac{y_k}{\pi_k}, \tag{2.25}$$

其中，π_k 表示选中概率。对于放回式 PPS 抽样方案，所对应的 HH 估算公式为（Hansen and Hurwitz，1943）：

$$\hat{t}_{HH} = \frac{1}{n}\sum_{k=1}^{n} \frac{y_k}{p_k} = \frac{1}{n}(\hat{t}_1 + \cdots + \hat{t}_k + \cdots + \hat{t}_n), \tag{2.26}$$

其中，每一个 $\hat{t}_k = y_k/p_k$ 估算总和 T。比率 R 的估计值 $\hat{r}$ 可以推导为两个 HT 估计值或是两个 HH 估计值的比率。另外，在估算中位值 M 时，将选中概率的倒数 $1/\pi_k$ 当成元素权重，来构建实际中的累积分布函数。

放回式假设也简化了设计方差的估算。对于总和的估计值 $\hat{t}_{HH}$，放回式 PPS 下的设计方差是：

$$V_{ppswr}(\hat{t}_{HH}) = \frac{N^2}{n}\sum_{k=1}^{N} p_k\left(\frac{Y_k}{Np_k} - \bar{Y}\right)^2 = \frac{1}{n}\sum_{k=1}^{N} p_k(T_k - T)^2, \tag{2.27}$$

其中，$T_k = Y_k/p_k$。从式 2.27 中可以推出，当 Y_k 与 Z_k 严格成比例，$Y_k/Z_k = C$ 时，设计方差为 0——在实际中几乎不存在的理想情形。方差的无偏估计为：

$$\hat{v}_{ppswr}(\hat{t}_{HH}) = \frac{N^2}{n(n-1)}\sum_{k=1}^{n}\left(\frac{y_k}{Np_k} - \bar{y}\right)^2 = \frac{1}{n(n-1)}\sum_{k=1}^{n}(\hat{t}_k - \hat{t}_{HH})^2 \tag{2.28}$$

其中，$\bar{Y}$与$\bar{y}$分别是研究变量 y 的总体均值与样本均值。

在系统 PPS 抽样中，以上估计值是一个近似值。无放回式以及拉奥-哈特

利-科克伦方法的近似方差估计值也可以推导出来。我们省略这些,读者可以参阅沃尔特(Wolter,1985)。

范例2.6

系统PPS抽样中的估算。以家庭数HOU85作为规模变量 z,使用PPSSYS,从1991年省级人口数据中抽取含有8个自治市的样本($n=8$)。总体的累积和为 $T_z=91\ 753$,PPSSYS中的抽样间隔是 $q=91\ 753/8 = 11\ 469$。

最大的单一元素"Jyväskylä"变量HOU85的取值是26 881。因此,元素"Jyväskylä"将被抽中两次,余下的6个样本元素将从剩下的总体元素(31个)中抽取。这种情形通常用下面的方式来解决。规模大于抽样间隔的元素一定会被抽中(但只抽中一次)。对于这一确定的元素,其权重和选中概率由定义而定为1。这里,我们首先将"Jyväskylä"放入第一个层级,并一定抽取。接下来,使用系统PPS方法,从第二个层级中的31个总体元素中抽取7个。这样,得到了含有8个自治市($n=8$)的样本。注意,表2.8中的样本根据规模变量HOU85调整过了。

表2.8 1991年省级人口中一个系统PPS样本($n=8$)

样本设计标识			元素标签	规模指标	研究变量	
STR	CLU	WGHT		HOU85	UE91	LAB91
1	1	1.000	Jyväskylä	26 881	4 123	33 786
2	10	1.004	Jyväsk. mlk.	9 230	1 623	13 727
2	4	1.893	Keuruu	4 896	760	5 919
2	7	2.173	Äänekoski	4 264	767	5 823
2	32	2.971	Viitasaari	3 119	568	4 011
2	26	4.762	Pihtipudas	1 946	331	2 543
2	18	6.335	Kuhmoinen	1 463	187	1 448
2	13	13.730	Kinnula	675	129	927

抽样比例(未用)。

在系统PPS设计中,建立一个权重变量很重要。对于总体元素 k,其权重 w_k 可以由下面公式计算得出:

$$w_k = \frac{1}{p_k \times n} = \frac{91\ 753}{Z_k \times n},$$

其中,Z_k 是元素 k 的HOU85取值。但是,这里的元素"Jyväskylä"确定会被抽中,其权重为1。余下层级二中的7个自治市元素的权重可以计算出来:

$$w_k = \frac{1}{p_k \times n} = \frac{91\ 753 - 26\ 881}{Z_k \times 7}。$$

在估算中,另一个需要的设计记号是层级标签STR。其取值为,确定的元素为1,其余为2。由于每个元素本身被当成一个独立的整群,元素编码被用作

CLU。另外,固定总体的校正因子 $(1-\sum_{k=1}^{n}p_k)$ 也可以利用,使得抽样与无放回式类型相似。表 2.9 中的估计值对应的是变量 UE91 的总和 $\hat{t}_{HT}$、比率 $\hat{r}_{HT}$ 与中位值 $\hat{m}_{HT}$。为了便于比较,表中也给出了总体参数 T,R 和 M 的数值。

表 2.9 1991 年省级人口中 PPSSYS 设计($n=8$)的估计值

统计量	变 量	参 数	估计值	s. e	c. v	deff
总和	UE91	15 098	15 077	521	0.03	0.0035
比率	UE91,LAB91	12.65%	12.85%	0.2%	0.02	0.1854
中位值	UE91	229	134	188	1.401	0.92

正如所期望的,PPSSYS 在估算总和中非常有效。$\hat{t}_{HT}$的设计效应估计值接近于 0(deff = 0.004)。这既是由于规模变量 HOU85 与研究变量 UE91 间的强相关引起的,也是因为这一估计值本身的线性特征造成的。对于非线性的比率估计值 $\hat{r}_{HT}$,PPSSYS 也相当有效,但比总和差很多。对于抗扰估计值 $\hat{m}_{HT}$,当前设计略微强于简单随机抽样。这部分是由于倾向于抽取大数目元素的 PPS 的特征。这些元素代表的是 UE91 分布的的边缘而非中间部分。

PPS 抽样的效率

我们使用总和 T 的估算过程来更详细地讨论 PPS 抽样的效率。估计值 $\hat{t}_{HT}$ 的 PPS 设计方差 $\hat{v}_{pps}(\hat{t}_{HT})$ 和固定总体中规模变量 z 与研究变量 y 的回归相关联。

$$Y_k = A + BZ_k + E_k,$$

其中,$E_k(k=1,\cdots,N)$ 是残差。

残差平方和与总体方差的关系由下面的公式给出:

$$\frac{1}{N-1}\times\sum_{k=1}^{N}(Y_k - A - BZ_k)^2 \approx S^2(1-\rho_{yz}^2),$$

其中,S^2 是 y 的总体方差,ρ_{yz}^2是变量 y 与 z 间的相关系数的平方。当这一相关系数接近于 ±1 时,残差离异很小。事实上,这一方差与后面要讨论的回归估计相同。因此,PPS 抽样的效率应该在上述回归模型的基础上来考察。但是,正如即将清楚的,只有强相关系数 ρ_{yz}并不能保证有效的估算。

探求 PPS 抽样的效率的一个简便的情形是,通过比较总和估计值在 SRSWR 与 PPSWR 中的方差,可以得到:

$$V_{srswr}(\hat{t}) - V_{ppswr}(\hat{t}_{HT}) = N^2\mathrm{Cov}\left(z,\frac{y^2}{z}\right)/n。$$

因此,当变量组$(z,y^2/z)$之间的相关系数为正时,PPS 抽样较 SRS 更为有效。另一方面,前面已经提到,当比率 Y_k/Z_k 对于每一个总体元素为常数——比如 C——时,PPS 抽样的效率最高。如果我们在上述协方差部分加入 $C=Y_k/Z_k$,$\mathrm{Cov}(z,y^2/z)$变成了 z 与 y 的协方差。因此,这里 z 与 y^2/z 之间的相关系

数等于原来变量 z 与 y 之间的相关系数。所以，我们得出结论，PPS 抽样比 SYSWR 更为有效的必要条件是，研究变量 y 与辅助变量 z 之间在总体中是正相关的。但充分条件是，比率 Y_k/Z_k 在总体中保持为常数。下面的例子将进一步来考察这两个条件。

范例 2.7

1991 年省级人口数据中 PPS 抽样的效率。为了评估 PPS 抽样的有效性，应当考察两个条件。它们是：比率 Y_k/Z_k 在总体中的稳定性，以及如果效率高时，回归直线 $\hat{Y}_k = 26.657 + 0.155 \times Z_k$ 与 y 轴的相交点应当接近原点。为了这些目的，从 1991 年省级人口数据中得出两个散点图，也计算了相应的系数。

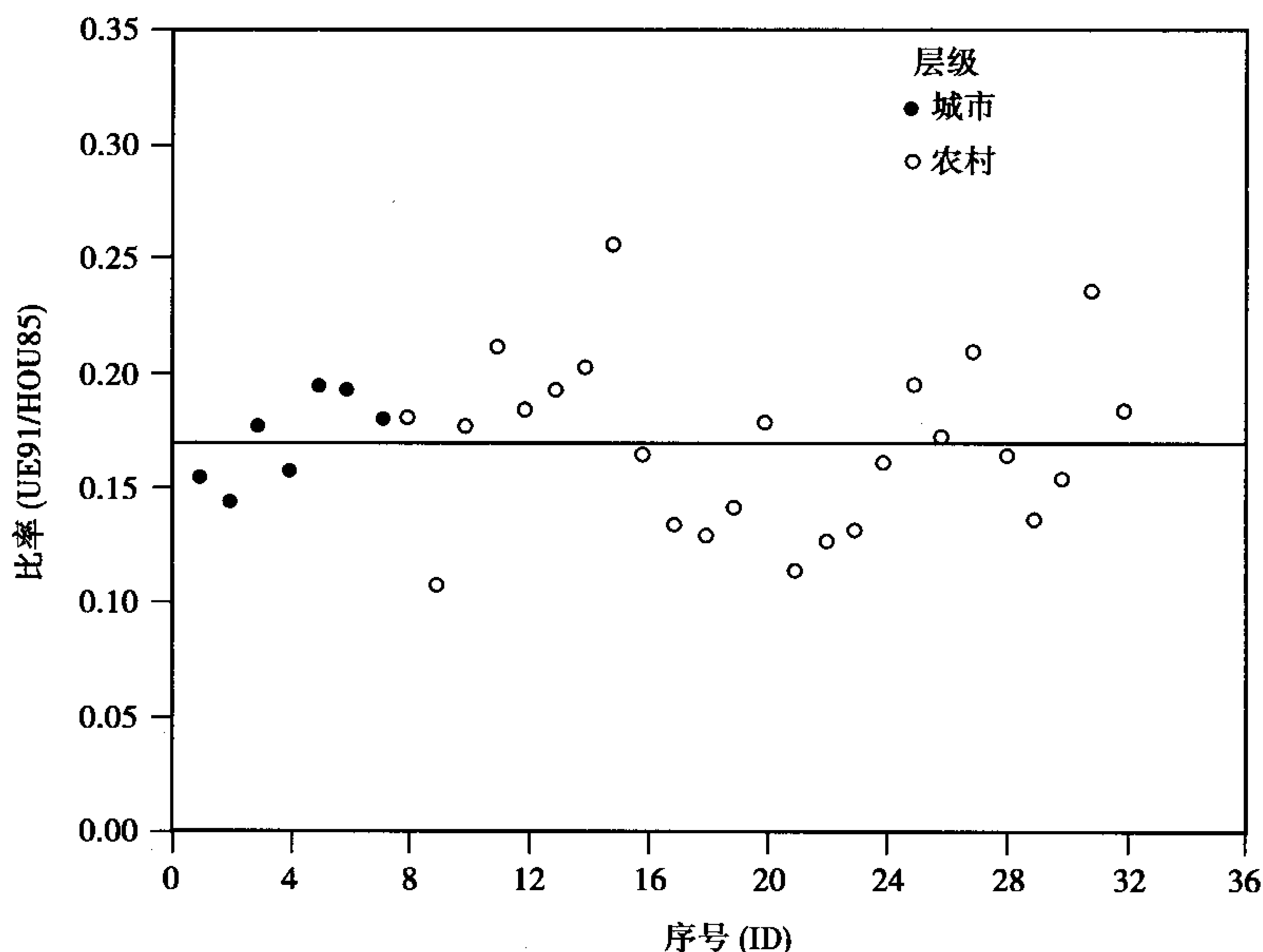


图 2.5　1991 年省级人口中比率 UE91/HOU85 对序号(ID)的散点图

图 2.5 给出了比率 Y_k/Z_k 在总体中的变动范围。正如这里的情形，这一比率接近于某一常数时，PPS 抽样比较有效。可以看出，最大的城镇在图的最左边(ID ≤ 7)。特别是这些城镇自治市中，比率 Y_k/Z_k 接近于一个常数。在 PPS 抽样中，倾向于抽中最大的元素，导致有效的总和的估计值。当比率 Y_k/Z_k 与比率 X_k/Z_k 为常数时，比率 Y_k/X_k 也会有同样的特征。

y 与 z 间的相关系数 $\rho_{yz} = 0.997$(参见图 2.6)。但是，强相关并不是 PPS 抽样有效估算的充分条件。让我们考虑一个完全相关的极端的例子，即，回归方程 $Y_k = A + B \times Z_k$ 完全吻合。使用回归系数的通常解释，当 A 较大时，回归直线与 y 轴的交点远离原点。因此，SRSWR 比 PPS 更为有效。从图 2.6 可以看出，在 1991 年省级人口数据中，家庭数 HOU85 解释了 99% 的失业人数

UE91 的离差，系数 A 近似为 0。

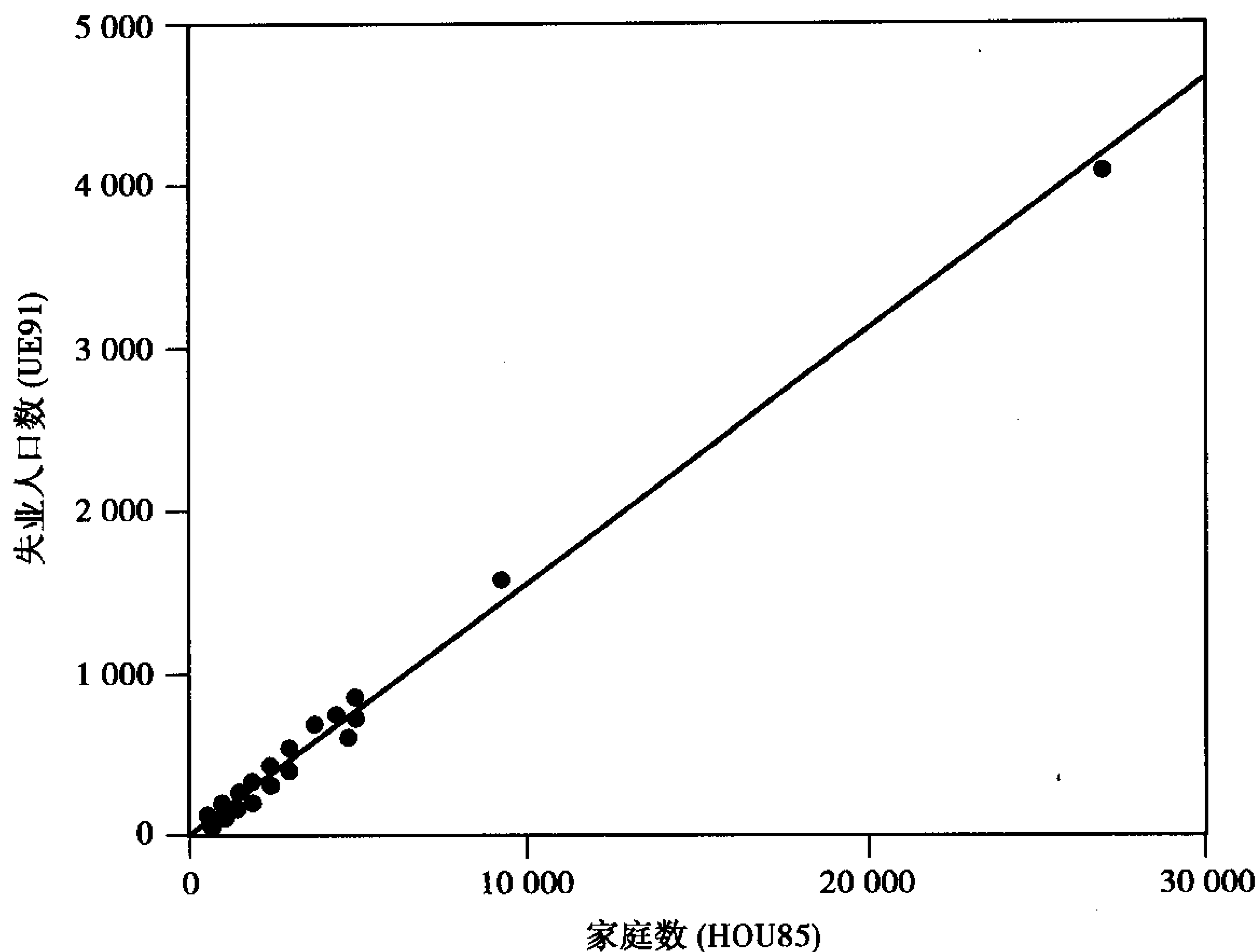


图 2.6 UE91 对 HOU85 散点图(1991 年省级人口)

小 结

从研究变量的取值范围较大的总体中抽取样本时，PPS 抽样提供了一项实用的技术，同时它经常会取得较高的效率。PPS 抽样的效率取决于两个方面。第一，效率根据所要估算的参数类型而变化。这里的参数是总和、比率以及中位值。对总和的估算看起来最为有效。在 PPS 抽样中，需要一个辅助规模变量(z)，并且，为了有效的估算，规模变量应当与研究变量 y 高度相关。这一条件即是变量组($z,y^2/z$)之间为正相关。在 1991 年省级人口数据中，这一条件是满足了的。但是，只满足这一条件并不能保证高效的估算。还需要比率 Y_k/Z_k 在总体中接近于一个常数。因为，这一条件在 1991 年省级人口数据中也得到了满足，PPS 提供了总和的高效估算。对于 PPS 抽样有更多兴趣的读者，可以参阅布鲁尔与哈尼福(Brewer and Hanif,1983)以及赫德亚特与辛纳(Hedayat and Sinha,1991,第 5 章)的书。

第3章 辅助信息的进一步使用

Further Use of Auxiliary Information

从总体元素中得来的辅助信息,可以用来设计易操作又高效的抽样方案,并在抽取样本后,进一步提高估计值的效率。根据各种关于总体框架中排列顺序的假设,我们在前面系统抽样中,使用辅助信息来选择适当的方差估算公式。在概率成比例的抽样(PPS)中,我们在样本选取阶段使用了辅助信息;恰当地选择辅助规模变量可以极大地提高效率。在章节3.1和3.2中,辅助信息将被用于分层抽样(STR)与整群抽样(CLU)。在这两种抽样技术中,辅助信息被用来设计抽样方案。在分层抽样中,主要的目的是提高效率;在整群抽样中,抽样和收集数据的实际需要是使用辅助信息的主要动机。

在使用于抽样设计以外,辅助信息还可用来提高已抽取的样本的估算效率。后续分层——样本抽取后的分层——可以使用定类辅助变量。如果有一个连续变量,并且它与研究变量高度相关,可以通过使用变量估算或是回归估算来提高效率。在这些方法中,统计模型将辅助信息引入到估算过程中。章节3.3将介绍这些模型辅助技术。这些技术的使用可以极大地提高估计值的准确性,即,得出接近于相应总体取值的估计值,并且降低这些估计值的设计方差。本书的扩展网页演示了这些。

分层抽样中的辅助信息

在分层抽样中,目标总体被分成称为层级的互不交叉的次级总体。它们被看成独立的总体。抽样就在各个总体中独立完成。为了分层,在抽样框中需要恰当的辅助信息。地域、人口和社会经济变量经常被当成分层的辅助变量。分层可以提高效率。这是因为,对于研究变量取值的期望变动范围而言,层级的构成通常使得相似的总体元素聚集在某一层级内。所以,层级内差异较小。

分层的信息有时是总体所固有的。比如,由于国家被分成互不交叉的地域性行政片区,层级可以被清楚地分辨出来。从各个区域中独立抽样,可以保证在样本中有整个国家各个地方的合适代表。这种行政类型的辅助信息可用

来设计抽样。分层也可以用于次级总体或是感兴趣组群的估算。可以定义重要的组群为层级，这样使得每一组群中能够分配到期望的样本规模（参见第 6 章）。另外，地域间或是层级间也可以进行比较。因此，除了成为一个创造内部同质的次级总体的工具以外，分层还可以成为估算过程和检验过程中的分类变量。

整群抽样中的辅助信息

与直接从总体元素中抽取样本不同，整群抽样从叫做整群的、自然生成的次级总体中抽取样本。比如，次级群体在实际中通常是公司里职员的集合、学校里学生的集合以及家庭中成员的集合。为了抽样，需要总体整群的名单；但是，并不需要总体中所有元素的完全名单，而仅仅是那些抽中的整群中的元素。对总体的了解，可以找出主要抽样单位。教育调查是一个使用这种结构的较好的例子。它的主要抽样单位通常是学校，首先要从学校名单中抽取一个学校样本。另外，抽样前，可以将总体整群分层。因此，整群抽样中的辅助信息，不仅仅是有关总体元素聚合成整群，还包括分层所需的整群的特征。

总体元素生成整群时，一群元素聚合在一起。这样的聚合通常倾向于以整群方式，它们在与调查相关的各个方面都很相似。群内同质将降低估算效率。但是，由于降低了实地花费，整群抽样可以是费用高效的。群内同质涉及叫做群内相关的实际参数。主要有两种方法恰当地考虑到了有效估算所需的群内相关。首先，群内相关可以被当成估算中的干扰效应，并以在估算和检验中剔除这样的干扰效应为目的。其次，整群的特征是可以被当成总体的结构现象而用模型来加以分析的。这样的话，总体可以被看成是一个等级或多层结构。例如在教育调查中，第一层是学校，第二层是班级，而第三层或是最低层级是学生。学生的成绩受到这种等级结构的制约。使用多层结构的模型分析法与这种方法相同，并且也假定相应的信息存在于数据之中。干扰方法与多层方法将分别在第 8 章和章节 9.4 中加以讨论。

估算过程中的辅助信息

通过使用章节 3.3 中讨论的模型辅助估算技术，辅助信息可以用来提高已有样本的效率。在模型辅助估算中，统计模型将辅助数据纳入到估算之中。在后续分层中，假定使用方差的线性分析或是 ANOVA 模型，其辅助数据由总体的各个部分及一个或多个定类变量的边缘频次组成。比率估算使用没有截距的线性回归模型，其辅助数据由一个或多个连续变量的总体总和构成，这些可以从官方统计数字中找到。在回归估算中，标准回归模型将辅助数据纳入估算过程中。这样的方法是通用回归估算（GREG）的特例。当研究变量与辅助变量间存在一种诸如高度相关的关系时，所有这些方法都可能比简单随机

抽样的估算更为有效。

3.1 分层抽样

将总体分层,使之成为互不交叉的次级总体,是另一种使用辅助信息来提高效率的流行技术。这样的辅助信息,通常在抽样框架的登记册或数据之中。典型的使用于分层的变量包括普查中收集的地区(如国家)、人口(如性别、年龄分组)以及社会经济(如收入分组)等变量。为了更好地利用分层抽样的高效,不仅要小心选择分层变量,还应当恰当地在各层级中分配样本数目。

分层抽样的流行有以下几个原因:

1. 由于行政管理上的因素,许多人口总体现有的自然划分可以用于分层。
2. 分层可以在抽样和估算过程中按层级使用辅助信息。
3. 如果层级内同质性高,分层可以增加估算的精度。
4. 分层可以保证所需要的小规模次级总体或是组群的代表性。

估算和设计效应

在分层抽样中,使用辅助信息将总体分成 H 个互不交叉的次级总体。它们的规模分别为 $N_1, N_2, \cdots, N_h, \cdots, N_H$,其总和为 N。从各个层级中独立地抽取元素,组成一个样本。各个层级中的样本数为 $n_1, \cdots, n_h, \cdots, n_H$。在分层抽样中,估计值通常是加权的层级估计值。各层的权重为 $W_h = N_h/N$。层级被看成是相互独立的次级总体。总和 T 的估计值 $\hat{t}$ 的公式如下:

$$\hat{t} = N\sum_{h=1}^{H} W_h \bar{y}_h = \sum_{h=1}^{H} \hat{t}_h = \hat{t}_1 + \cdots + \hat{t}_h + \cdots + \hat{t}_H, \tag{3.1}$$

其中,$\hat{t}_h = N_h \bar{y}$ 是层级 h 中的总和估计值,而 $\bar{y}_h = \sum_{k=1}^{n_h} y_k/n_h$。当所有层级的总和估计值是无偏估计时,总体总和也是无偏的。因为样本是从各层级独立抽取的,$\hat{t}$ 的设计方差 $V_{str}(\hat{t}_{str})$ 即是层级方差 $V(\hat{t}_h)$ 的和。例如,如果各层级中使用了无放回式简单随机抽样,估计值 $\hat{t}$ 的设计方差是:

$$V_{str}(\hat{t}) = \sum_{h=1}^{H} V_{srs}(\hat{t}_h), \tag{3.2}$$

其相应的无偏估计值为:

$$\hat{v}_{str}(\hat{t}) = \sum_{h=1}^{H} \hat{v}_{srs}(\hat{t}_h)。 \tag{3.3}$$

$\hat{t}$ 的设计效应(DEFF)在很大程度上取决于由层级间与层级内方差造成的总体离异的比例。从等式 3.2 中可以推出,为了获得较小的设计方差,应当建构内部同质性高、层内方差小的层级。由于个体层级方差也取决于各个层

级的规模大小，这一效率同时受制于分配方案。让我们来计算总和 T 的 DEFF。我们使用了层级规模为 $n_h = n \times W_h$，而 $n = \sum_{h=1}^{H} n_h$ 的按比例分层的抽样技术。当各层级内的元素使用了无放回式简单随机抽样（SRSWOR），$\hat{t}$ 是 T 的无偏估计，而 $\hat{t}$ 的设计方差是：

$$V_{str}(\hat{t}) = N^2\left(1 - \frac{n}{N}\right)\sum_{h=1}^{H} W_h \frac{S_h^2}{n},$$

其中，S_h^2 是 y 在层级 h 中的方差。另外，$\hat{t} = (N/n)\sum_{h=1}^{H} y_k$ 的 SRSWOR 方差 $V_{srs}(\hat{t}) = N^2(1 - n/N)S^2/n$ 可以被写成如下的分层抽样的形式。给定大数 n，我们有：

$$V_{srs}(\hat{t}) \approx N^2\left(1 - \frac{n}{N}\right)\frac{\left[\sum_{h=1}^{H} W_h S_h^2 + \sum_{h=1}^{H} W_h(\bar{Y}_h - \bar{Y})^2\right]}{n},$$

其中，$\bar{Y}_h$ 是层级 h 中的总体均值，中括弧中的第一项表示层内离异，平方差 $(\bar{Y}_h - \bar{Y})^2$ 表示层级均值与总体均值 $\bar{Y}$ 间的离异，即层间离异。总体方差被分成层内与层间方差两部分，所以，$\hat{t}$ 的 DEFF 为：

$$\mathrm{DEFF}_{str}(\hat{t}) \approx \frac{\sum_{h=1}^{H} W_h S_h^2}{\sum_{h=1}^{H} W_h[S_h^2 + (\bar{Y}_h - \bar{Y})^2]}, \tag{3.4}$$

或是使用方差分析类似的形式：

$$\mathrm{DEFF}_{str}(\hat{t}) \approx \frac{\text{层内方差}}{\text{总体方差}} = \frac{\mathrm{MSW}}{S^2},$$

其中，总体方差 = 层内方差 + 层间方差。

范例 3.1

接下来，我们来计算从 1991 年省级人口数据中按比例配额的简单随机分层抽样的参数 $\mathrm{DEFF}_{str,pro}(\hat{t})$。总体由两个部分组成：层级 1 是城镇自治市（$N_1 = 7$），层级 2 是农村自治市（$N_2 = 25$）。把这两个层级当成 ANOVA 中某一因子的两个取值，我们在表 3.1 中得到研究变量 UE91 总体离异的分解。带入式 3.4 中的层内方差部分 $MSW = 4.35 \times 10^5$ 以及总体方差 $S^2 = 5.53 \times 10^5$，有：

$$\mathrm{DEFF}_{str,pro}(\hat{t}) \approx \frac{4.35}{5.53} = 0.79,$$

表 3.1 两层级（$H = 2; N_1 = 7, N_2 = 25$）分层 SRSWOR 抽样的总体 ANOVA 表

离异来源	df	平方和	平均平方和
层间	1	$SSB = 40.73 \times 10^5$	$MSB = 40.73 \times 10^5$
层内	30	$SSW = 130.60 \times 10^5$	$MSW = 4.35 \times 10^5$
合计	31	$SST = 171.32 \times 10^5$	$S^2 = 5.53 \times 10^5 = 743^2$

它与 DEFF 的实际计算值 $DEFF_{str,pro}(\hat{t})=0.84$ 相当接近。

按比例分配样本提供了一个简单的配额方法。看起来，按比例分配的简单随机分层抽样比 SRSWOR 更有效。下面，我们将讨论效率更高的其他配额方法。为了达到这一目的，应当更为有效地考虑各层级方差。

样本配额

给定样本的总数为 n 的限制条件，配额是一种可以确定每一个层级中抽取多少元素的工具。在一个相对受限的只有一个研究变量的描述性调查中，谨慎的目标是找到一个使得估算高效的分配方案。但是，应当注意到，在大规模的分析调查中，要达到分层抽样整体最佳的分配方案是不可能的。这是因为，这样的调查中研究变量太多。

配额的最佳与否取决于层级规模。它更多地取决于研究变量的总体方差在层间与层内的份额。在文献的大量配额方法中，除了按比例配额外，我们还将讨论最佳或是内曼配额与指数或是班基尔配额。

1. 按比例配额。这是最简单的分配方案，并且在实际中广泛使用。由于抽样比例 n_h/N_h 在各个层级中是常数，它仅需要预先知道层级规模。层级 h 中的样本数 n_h 是：

$$n_{h,pro}=n\times\frac{N_h}{N}=n\times W_h,$$

其中，W_h 是层级权重。

按比例配额保证了各层级中的抽样数相等，但可能生成较普遍预期低效的估计值。

因为，各层级的抽样比例是 n/N，总体中任一元素 k 的选中概率也为常数 $\pi_k=\pi=n/N$。这一方案因而是与 SRSWOR 相同的等概抽样设计。这一特征简化了估算过程，因为

$$\hat{t}=N\sum_{h=1}^{H}\sum_{k=1}^{n_h}\frac{y_{hk}}{n},$$

而不用计算层内均值。正是这一原因，按比例配额抽样有自我加权的特征。这样的特征，在其他选中概率依层级变化的配额方案中是没有的。

2. 最佳或是内曼配额。当各层中研究变量的标准差 S_h 已知时，可以使用这一方法。层级 h 中的样本数 n_h 是：

$$n_{h,opt}=n\frac{N_hS_h}{\sum_{h=1}^{H}N_hS_h}。$$

在实际中，很少知道 S_h。但从过去调查的经验中可以得到标准差真实值的近似值。在最佳配额中，与规模小或是内部同质性高的层级相比，

规模大或是标准差大的层级能提供更多的样本。在分层抽样中,这种配额的估计值最为高效。

3. 指数或是班基尔配额。这种方法使用于有着数量众多的小层级,并且需要在各层级中作出精确估算的情形。比如,在指数配额中,需要达到有效估算层级总和的 n_h 为:

$$n_{h,pow} = n\frac{(T_{hz})^a C.V_{hy}}{\sum_{h=1}^{H}(T_{hz})^a C.V_{hy}},$$

其中,T_{hz}是辅助变量 z 的层内总和,$C.V_{hy}$是 y 在层级 h 中的离异系数(C.V)。常数 a 被称为配额的指数。在实际中,a 恰当的取值可以是 1/2 或是 1/3。这样的选择可以被看成是内曼配额与各层级配额接近于常数间的妥协。

范例 3.2

1991 年省级人口数据中不同配额方案的简单随机分层抽样。首先,总体被分成两个层级,一为城镇,另一为农村。在所有自治市中,7 个($N_1=7$)为城镇,其余为农村地区($N_2=25$)。简单随机分层抽样抽取了 8($n=8$)个自治市。在按比例配额、最佳配额及指数配额的方案中,我们计算了大致的层级规模。层级的某些背景信息在表 3.2 中。

表 3.2　1991 年省级人口中变量 UE91 的层级参数

统计量	层级 1	层级 2	合　计
均值	1 146	283	472
总和	8 022	7 076	15 098
标准差	1 318	331	743
离异系数	1.150	1.170	1.572
层级规模	7	25	32

从表 3.2 中,在各种配额方案下的各个层级的 n_h 的计算值为,

(1)按比例配额:

$$n_{h,pro} = n\frac{N_h}{N} = \begin{cases} n_1 = 8 \times \frac{7}{32} = 1.75 \\ n_2 = 8 \times \frac{25}{32} = 6.25 \end{cases}$$

(2)最佳配额:

$$n_{h,opt} = n\frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h} = \begin{cases} n_1 = 8 \times 9\,226 \div (9\,226 + 8\,275) = 4.22 \\ n_2 = 8 \times 8\,275 \div (9\,226 + 8\,275) = 3.78 \end{cases}$$

(3)指数配额(近似;$a=0$):

$$n_{h,a=0} = n \times \frac{\mathrm{C.V}_{hy}}{\sum_{h=1}^{H}\mathrm{C.V}_{hy}} = \begin{cases} n_1 = 8 \times \dfrac{1.150}{1.150+1.170} = 3.97 \\ n_2 = 8 \times \dfrac{1.170}{1.150+1.170} = 4.03 \end{cases}$$

(3′)指数配额(准确;$a=0$;层级分组系数 c_h):

$$n_{h,a=0} = n \times \frac{\mathrm{C.V}_{hy}}{\sum_{h=1}^{H}\mathrm{C.V}_{hy}} \times c_h = \begin{cases} n_1 = 8 \times \dfrac{1.150}{1.150+1.170} \times 0.81 = 3.22 \\ n_2 = 8 \times \dfrac{1.170}{1.150+1.170} \times 1.19 = 4.78 \end{cases}$$

这些计算导出以下结论。在按比例配额中,各个层级的样本规模为 $n_1=2$ 和 $n_2=6$。在最佳配额与近似指数配额中,$n_1=4$ 和 $n_2=4$。注意这里的等额也给出了 $n_1=n_2=4$,它表示各个层级中的样本数目相等($n_h=n/H$)。根据各个方案相应的 DEFF 值(最佳配额 0.44;指数配额 0.51;按比例配额 0.74)可以推出,前两者较后者更高效。

另外,由于这里的小总体情形(比如,1991 年省级人口数据)并不能满足每层级低抽样比例的假设,我们也计算了准确指数配额的各个估计值。当 $a=0$ 时的准确指数配额方案下的抽样规模为 $n_1=3$ 和 $n_2=5$。

我们也可以通过计算离异系数 $\mathrm{C.V}(\hat{t})$ 或是样本总体与各层级的相对标准误来比较各配额方案。表 3.3 给出了这些结果。在总体层次上,正如我们所期望的,最佳配额的估计值最精确,$\mathrm{C.V}(\hat{t})=0.32$。但是,在层级的层次上,准确指数配额的估计最为精确,因为其 $\mathrm{C.V}(\hat{t})$ 在两个层级上均接近 0.5。按比例配额在总体层次上的精确度较差,同时其在两个层级上的离异系数差距较大。

表 3.3 各种配额方案下层级样本规模与离异系数
(1991 年省级人口 STRSRS 样本中总和的估算)

配额方案	样本规模		层级		总体	
	n_1	n_2	$\mathrm{C.V}(\hat{t}_1)$	$\mathrm{C.V}(\hat{t}_2)$	$\mathrm{C.V}(\hat{t})$	DEFF
最佳	4	4	0.38	0.54	0.32	0.44
指数(准确)	3	5	0.50	0.47	0.35	0.51
按比例	2	6	0.68	0.42	0.42	0.74

前面提到过,在样本配额前,当组群与层级重合时,使用指数配额(近似或是准确)会大大提高精确度。第 6 章将详细讨论组群估算过程。

样本选取

在各个层级中,样本的选取独立进行。这使得可以在不同的层级中使用不同的抽样方案。但是,为了方便,通常使用相同的抽样方案。在 STR 抽样中,首先将总体分层,然后在各层级内抽取随机样本。在层级层次上,简单随

机抽样、SYS 或是 PPS 都可以使用。

选中概率取决于层级样本选取的方法。比如，在所有层级中使用 SRSWOR，得出 $\pi_{hk}=n_h/N_h$，其中，n_h 是层级样本规模，而 N_h 是层级 h 中总体元素的总数。如果使用 PPS 抽样，则选中概率为 $\pi_{hk}=n_h\times(Z_{hk}/T_{hz})$，其中，$T_{hz}$ 是规模变量 z 的层级总数 $T_{hz}=\sum_{k=1}^{N_h}z_{hk}$。我们需要选中概率来定义相应的抽样权重。接下来，我们将讨论 1991 年省级人口数据中的最佳配额分层抽样。

范例 3.3

1991 年省级人口数据的最佳配额简单随机分层抽样。演示总体分成两个层级——农村与城镇自治市。配额方案是最佳配额方法，因而在估算总体总和 T 时，各层级的样本规模相等 $n_1=n_2=4$。在这样的配额方案下，抽取了简单随机分层抽样样本(表 3.4)。抽取样本过后，需要将相关的设计标签作为新变量(STR，CLU 和 WEIGHT)加入数据之中，并应用于估算过程中。与前面相同，有 3 个估算问题。总失业人数 UE91 的估计值 $\hat{t}_{str}$ 清楚地演示了，分层降低了标准误(表 3.5；原书为表 3.4。——译者注)。在失业比率 UE91/LAB91 的估计值 $\hat{r}$ 中，我们发现了相似的效应。第三个 UE91 总体分布中位值的估计值 $\hat{m}$ 中，使用分层和最佳配额没有提高结果。

表 3.4 1991 年省级人口的一个最佳配额分层的简单随机样本

抽样设计标识			元素标签	研究变量	
STR	CLU	WGHT		UE91	LAB91
1	1	1.75	Jyväskylä	4 123	33 786
1	2	1.75	Jämsä	666	6 016
1	4	1.75	Keuruu	760	5 919
1	6	1.75	Suolahti	457	3 022
2	21	6.25	Leivonmäki	61	573
2	25	6.25	Petäjävesi	262	1 737
2	26	6.25	Pihtipudas	331	2 543
2	27	6.25	Pylkönmäki	98	545

抽样比率：层级 1 =4/7 =0.57；层级 2 =4/25 =0.16。

层级标签的取值是，城镇 STR =1，农村 STR =2。整群标签 CLU 表示元素的集合；这里每一个整群中只有一个元素，因而自治市的编号成为它们的整群标签。权重变量需要根据各层中的层级规模以及层级样本数目来计算。权重变量 WEIGHT 的取值为，第一层 $w_{1k}=N_1/n_1=7/4=1.75$，第二层 $w_{2k}=N_2/n_2=25/4=6.25$。另外，在无放回式简单随机抽样中，每一层级的抽样比例应该给出来，第一层为 4/7 =0.57，第二层为 4/25 =0.16。

表 3.5 1991 年省级人口一个最佳配额分层的简单随机样本的估计值

统计量	变 量	参 数	估计值	s. e	c. v	deff
总和	UE91	15 098	15 211	4 286	0.28	0.21
比率	UE91,LAB91	12.65%	12.78%	0.3%	0.02	0.38
中位值	UE91	229	177	64	0.36	0.19

表 3.5 给出了估算结果以及相应的总体参数。总和点估计 $\hat{t}$ 和比率点估计 $\hat{r}$ 与总体参数 T 和 R 相当接近。但是,中位值点估计 $\hat{m}=177$ 与真实中位值 $M=229$ 相去甚远。这里,最佳配额的简单随机分层抽样的设计对于估算总和与比率非常高效。设计效应的估计值 $\text{deff}(\hat{t})=0.21$,以及 $\text{deff}(\hat{r})=0.38$。但是,中位值的估算较没有分层的 SRSWOR 设计更为有效。因为 $\text{deff}(\hat{m})=0.19$,远小于 1。

最后,我们计算了变量 UE91 的总和估计值的层级精度 c. v。第一层的总和估计值 $\hat{t}_1=10\ 507$,第二层的总和估计值 $\hat{t}_2=4\ 700$。相应的标准误估计值为 $\text{s.e}(\hat{t}_1)=4\ 015$ 与 $\text{s.e}(\hat{t}_2)=1\ 481$。因而,c. v 估计值为 $\text{c.v}(\hat{t}_1)=0.38$ 与 $\text{c.v}(\hat{t}_2)=0.32$。两者相差不大。

小 结

我们讨论了将小规模总体分成两个部分并使用不同的配额方案。在估算总和、比率与中位值时,最佳配额的分层抽样有效地估算了总和与比率,生成了相应的 deff 估计值。但是,估算的中位值的 deff 大于 1。总的来讲,分层抽样中精确度的整体提高,取决于分层方案以及各层级间配额的分配。在层级中,特别是,对各个层级分别计算估计值时,精确度受到合适的配额方案的影响。

分层是一个提高效率的有力工具,并适用于各种情形。因而,在实践中,它得以广泛使用。除了抽取样本元素外,分层抽样也常见于抽取整群的复杂调查的抽样设计中。

3.2 整群抽样

在复杂调查中,自然形成的总体元素,诸如家庭户、村庄、城市街区或是学校,经常被用于抽样和数据收集中。比如,家庭户可以作为面谈调查的资料收集单位。除了原始的个人层次的总体以外,还有家庭户的总体。假定有合适的框架,可以在面谈样本中家庭成员之前,抽取一个家庭户的样本。这是一个一级整群抽样的例子。如果没有家庭户的总体框架,但是有街区层次的框架,

就可以先从街区登记中抽取街区，再为抽中的街区准备其中的居住家庭户的登记名单，从名单中抽取家庭户。这是一个两级整群抽样的例子。

在社会和商业调查中，使用整群抽样的动机是实际、经济以及行政上的效率。整群抽样的一个重要优势是，并不需要元素层次的抽样框架。需要的仅仅是，整群层次的框架以及抽中整群的元素框架。整群层次的框架通常很容易得到，比如公司、学校、街区或是街区类的单位。同时，这些已有的结构提供了把重要的结构信息纳入分析中的机会。比如，教育调查实践中，学生被划分到学校中，并进一步划分到学校内的班级或是年级中。一组学校可以被看成是整群（学校）的集合，从中抽取一个学校样本；接下来，从抽中的学校中抽取一个年级的样本。如果调查抽中年级所有的学生，则这样的设计是二级整群抽样。除了抽样和收集数据，多层结构也可以用于分析之中。比如，可以比较学校间的差异。

因此，在多级抽样中，除了最后一级外，可从各级抽中的整群中抽取一个次级样本。在最后一级中，抽中整群中的所有元素成为元素层次的样本，或是抽取一个次级元素样本。在本章中，我们讨论一级和二级整群抽样并使用 1991 年省级人口数据做演示。第 5 章到第 9 章使用各种真实的调查来进一步演示包括整群总体分层的整群抽样。

整群抽样的经济动机是收集样本元素的低成本。特别是对于区域分布广泛的总体，更是如此。由于访谈员的工作可以根据区域来计划，他们的旅行费用可以极大地降低。所以，整群抽样的费用效率很高。但是，整群抽样也有某些统计效率上的缺陷。如果整群完全反映出总体结构，我们可以有效的抽样，使得标准误的估计值小于简单随机抽样的估计值。但是，在实际中，整群倾向于内部高同质性。而这种同质性增加标准误，因而降低统计效率。我们将通过研究群内相关来讨论这一问题。这一概念将在后面的章节分析真实数据中得来的整群抽样中得以广泛使用。有两种方法来讨论它：将群内相关当成干扰效应与使用多级模型的方法。

整群抽样中的费用效率

让我们首先使用一个简单例子来说明整群抽样相对于 SRSWOR 的费用效率。整群抽样的费用效率可以用一个简单的费用函数来表示：

$$C_{clu} = c_1(m) + c_2(m \times B),$$

其中，C_{clu}——抽样总费用；

c_1——整群抽样费用；

c_2——抽取整群中一个元素的费用；

B——整群中元素的数目（等规模整群）；

m——整群样本数目；

$n = m \times B$——样本中元素的数目。

在 SRSWOR 的情形中,费用函数是:

$$C_{srs} = c_1 n + c_2 n,$$

其中,n 是样本中元素的数目。

同等总费用的限制条件下,得到以下 SRSWOR 和 CLU 抽样规模:

$$n_{srs} = \frac{C}{c_1 + c_2}$$

$$n_{clu} = \frac{C}{\left(\frac{1}{B}\right)c_1 + c_2},$$

由此表明,给定抽样费用,整群抽样比 SRSWOR 可以调查更多的总体元素。同时,标准误随着样本规模的增加而降低,这可以抵消群内同质性对于标准误的反作用。这意味着,DEFF 可以成为整群抽样总体效率的标尺,因为它也考虑到了费用效率。

范例 3.4

整群抽样的费用效率。全国范围内计算机辅助的个人访谈(CAPI)的预算包括抽样和收集数据的费用15 000 欧元。每人的访谈费用是30 欧元,每个访谈的平均旅费为35 欧元。首先,我们假定 SRSWOR 抽样方法,给定总费用的样本规模为:

$$n_{srs} = \frac{15\ 000}{35 + 30} = 231。$$

接下来,我们假定总体被分为整群,每个整群中含有 5 人($B = 5$),样本规模为:

$$n_{clu} = \frac{15\ 000}{\frac{35}{5} + 30} = 405。$$

由于每一份旅费可以访谈 5 人,整群抽样差不多是 SRSWOR 样本的两倍。

一级整群抽样

让我们介绍最简单的整群抽样设计原则,即是一级整群抽样。在一级整群抽样中,假设总体中的 N 个元素被分成 M 个小组,即整群。如果做出并不符合实情的假设,每个整群中含有 B 个元素。在更加普遍的情形下,整群 i 的元素数目为 B_i。在这两种情况下,从总体的 M 个整群中抽取 m 个整群,并且抽中整群的所有元素均纳入最终样本中。记住,这里只是一级抽样,即在整群层次。因而,这样的设计被称为一级整群抽样。

从总体整群中抽取这 m 个整群时,使用具体的元素抽样技术,如 SRS,SYS 或是 PPS 抽样。由于标准的元素抽样方案可以用于一级整群抽样,所以

前面讨论过的抽样技术是现成的。唯一的区别在于,含有一组总体元素的整群,而非总体中的单个元素,构成了抽样单位。同时,如果整群抽取中使用了等概的方法,比如 SRSWOR 或是 SYS,那么,总体元素的选中概率也是相等的。不管整群规模是否一致,都是如此。

在相等规模整群的简单例子中,元素样本的规模是给定的,$n = m \times B$。如果正如实际情况,整群规模不等,样本规模 $n = \sum_{i=1}^{m} B_i$ 并不能事先确定,它取决于样本最后抽取了那些整群。如果整群规模差异较大时,预期的元素样本 $(m/M) \times N$ 与实际样本规模 n 可能有较大差别。使用合适的抽样方案,可以控制这样的麻烦。比如,总体中的整群规模(即使是大致的)是已知的辅助信息(即使是粗略的),就可以将整群分层,以便大致控制元素样本的规模。

对于总体为 M 个规模为 B_i 的整群,使用了 SRSWOR 抽取 m 个整群,我们称之为一级 CLU 抽样,其中,等规模整群 $B_i = B$ 是它的特例。我们介绍其估算过程的基本情况。元素层次的总体规模是 $N = \sum_{i=1}^{M} B_i$。我们的目标是估计总体总和 T。为了达成这一目标,可以使用章节 2.3 中简单随机抽样的公式,并将它用于整群抽样。我们也将给出另外一些估算公式。

让总体中研究变量的取值写成 $Y_{ik}, i = 1, \cdots, M$,样本中的写成 $y_{ik}, i = 1, \cdots, m$,其中均有 $k = 1, \cdots, B_i$。总体中整群的总和 T_i 为:

$$T_i = \sum_{k=1}^{B_i} Y_{ik} = B_i \bar{Y}_i, \qquad i = 1, \cdots, M,$$

其中,$\bar{Y}_i$ 是总体中整群 i 里每一元素的均值,其样本估计值为 $\bar{y}_i = \sum_{k=1}^{B_i} y_{ik}/B_i$,$i = 1, \cdots, m$。

总体总和的 $T = \sum_{i=1}^{M} T_i$ 的无偏估计量是:

$$\hat{t} = (M/m) \sum_{i=1}^{m} B_i \bar{y}_i \text{。} \tag{3.5}$$

从相应的 SRSWOR 的公式中,可以推导出 $\hat{t}$ 的设计方差 $V_{clu-I}(\hat{t})$ 以及它的无偏估计值 $\hat{v}_{clu-I}(\hat{t})$。唯一的离异来源是整群总和 T_i 对于每一整群的总和均值 $\bar{T}_M = \sum_{i=1}^{M} T_i/M$。

$\hat{t}$ 的设计方差是:

$$V_{clu-I}(\hat{t}) = M^2(1 - m/M) \sum_{i=1}^{M} (T_i - \bar{T}_M)^2/m(M-1) \text{。} \tag{3.6}$$

这一设计方差的无偏估计量为:

$$\hat{v}_{clu-I}(\hat{t}) = M^2(1 - m/M) \sum_{i=1}^{m} (B_i \bar{y}_i - \hat{\bar{T}}_m)^2/m(m-1), \tag{3.7}$$

其中 $\hat{\bar{T}}_m = \sum_{i=1}^{m} B_i \bar{y}_i/m$ 是每整群均值 $\bar{T}_M$ 的估计值。

从式 3.6 中可以推出，当整群规模 B_i 相等或是接近于相等以及整群均值 $\overline{Y}_i$ 变化较小时，各整群总和 $T_i = B_i\overline{Y}_i$ 变化较小，因而可以得到一个较小的设计方差。另一方面，当整群规模的变化较大，则整群的总和差异较大，设计方差将变得较大。但是，使用一个比率估计量可以提高效率。其中，整群规模 B_i 被当成辅助规模变量 z。这样，我们得到总和的估计值如下：

$$\hat{t}_{rat} = N\frac{\sum_{i=1}^{m} T_i}{\sum_{i=1}^{m} B_i} = N \times \bar{y}, \tag{3.8}$$

其中，$\bar{y} = \sum_{i=1}^{m} T_i / \sum_{i=1}^{m} B_i$ 是样本的元素均值。它是总体的元素均值 $\overline{Y} = T/(M \times B)$ 的估计值。这一比率估计值是稍后章节 3.3 中讨论的比率估计值的一个特例。假定整群样本的数目较大，设计方差 $\hat{t}_{rat}$ 的近似值为：

$$V_{clu-I}(\hat{t}_{rat}) \approx M^2(1 - m/M) \sum_{i=1}^{M} B_i^2(\overline{Y}_i - \overline{Y})^2/m(M-1)。 \tag{3.9}$$

整群均值 $\overline{Y}_i$ 与总体均值 $\overline{Y}$ 间的离异，在预期上通常小于整群总和 T_i 与群间均值 $\overline{T}_M = \sum_{i=1}^{M} T_i/M$ 间的离异。如果是这样，估算过程将更加有效。所以，设计方差的公式为：

$$\hat{v}_{clu-I}(\hat{t}_{rat}) = M^2(1 - m/M) \sum_{i=1}^{m} B_i^2(\bar{y}_i - \bar{y})^2/m(m-1)。 \tag{3.10}$$

如果我们事先知道各整群的规模 B_i，使用 PPS 抽样方法可以有同样的提高效率的作用。所以，可以使用与章节 2.5 相应的 PPS 公式。

也可以通过整群均值 $\bar{y}_i$ 的均值 $\bar{y}_m$ 来估算总和 T：

$$\bar{y}_m = \sum_{i=1}^{m} \frac{\bar{y}_i}{m},$$

这是整群均值在总体上的均值 $\overline{Y}_M = \sum_{i=1}^{M} \overline{Y}_i/M$ 的估计值。当整群规模相等，即当 $B_i = B$ 时，得到估算公式：

$$\hat{t}_m = N\bar{y}_m = N\sum_{i=1}^{m} \bar{y}_i/m 。 \tag{3.11}$$

对于 T 而言，它是无偏的，并与式 3.5 中的 $\hat{t}$ 和式 3.8 中的 $\hat{t}_{rat}$ 是相等的。但是，这一 $\hat{t}_m$ 在整群规模不等时，可以是有偏甚至是不一致的。从偏差中可以进一步看清楚。偏差是：

$$\mathrm{BIAS}(\hat{t}_m) = -\sum_{i=1}^{M} (B_i - \overline{B})(\overline{Y}_i - \overline{Y}_M),$$

其中，$\overline{B}$ 是平均整群规模。这一等式表明，当整群规模 B_i 与整群均值 $\overline{Y}_i$ 并不相关时，估算值 $\hat{t}_m$ 是无偏的。当整群规模相等时，就会满足这样的条件。因此，要使用 $\hat{t}_m$ 的话，需要仔细检查整群规模与整群均值间的关系。

在整群规模相等的情形下,可以给出 $\hat{t}_m$ 的设计方差:

$$V_{clu-1}(\hat{t}_m) = (M \times B)^2(1 - m/M)S_b^2/m, \tag{3.12}$$

其中,群间方差 S_b^2 可以根据整群均值 $\bar{Y}_i$ 及其均值 $\bar{Y}_M$ 推算出来:

$$S_b^2 = \sum_{i=1}^{M}(\bar{Y}_i - \bar{Y}_M)^2/(M-1)。$$

由于整群规模相等,所以可以使用 $\hat{t}$ 与 $\bar{Y}$ 在式 3.12 与 S_b^2 中来代替 $\hat{t}_m$ 与 $\bar{Y}_M$。

接下来,我们将通过考察一级整群抽样设计下的总和估计值的 DEFF,来研究这样的设计的效率。其中,我们假定整群规模是相等的。

范例 3.5

1991 年省级人口数据中一级整群抽样的效率。我们将在失业人数(UE91)总和 T 的估算过程中,通过计算 T 估计值的 DEFF 来讨论一级整群抽样的效率。捆绑相邻的 4 个自治市,生成 8 个整群。全省 $N=32$ 个自治市被分成 $M=8$ 个规模相等的整群 $B_i=B=4$。应当注意到,在实际调查中,整群的规模通常是不等的,而且总体中的整群数目要大得多。因此,以下的计算仅仅是假设性的,用来演示一些估算的原则而已。表 3.6 给出了所有总体整群的整群均值 $\bar{Y}_i$ 与 UE91 的总和 $\bar{T}_i$。

整群总和 T_i 的合计等于总体总和 $T=15\ 098$。由于整群规模相等,总体中的元素均值与整群均值的均值 $\bar{Y}_M$ 都为 472。让样本规模为 $m=2$ 个整群,样本中的元素规模为 $n=m \times B=2 \times 4=8$。由于整群规模相等,$\hat{t}$,$\hat{t}_m$ 及 $\hat{t}_{rat}$的估计值相同,其中任一相应的设计方差均可以使用。为了评估效率,我们使用式 3.12 来计算 $\hat{t}$ 的设计方差。

表 3.6 1991 年省级人口数据中含 4 个相邻市的整群均值与总和

整群标识			人口整群中 UE91 的均值与总和	
STR	CLU	元素(含市)	均值 $\bar{Y}_i$	总和 T_i
1	1	Jyväskylä, Korpilahti, Muurame, Säynätsalo	1 206	4 824
1	2	Jämsä, Jämsänkoski, Keuruu, Kuhmoinen	535	2 141
1	3	Saarijärvi, Konginkangas, Äänekoski, Sumiainen	427	1 709
1	4	Kannonkoski, Karstula, Kyyjärvi, Pylkönmäki	172	686
1	5	Suolahti, Hankasalmi, Konnevesi, Laukaa	481	1 923
1	6	Joutsa, Leivonmäki, Luhanka, Toivakka	109	436
1	7	Jyväskylä mlk., Multia, Petäjävesi, Uruainen	556	2 223
1	8	Kinnula, Kivijärvi, Pihtipudas, Viitasaari	289	1 156

整群总和之和 $T=15\ 098$;

整群均值 $\bar{T}_M=1\ 887$;

元素均值 $\bar{Y}=472$;

整群均值之均值 $\bar{Y}_M=472$。

首先，得到群间方差为：

$$S_b^2 = \frac{1}{(8-1)} \sum_{i=1}^{8} (\bar{Y}_i - 472)^2 = 340^2,$$

其给出设计方差：

$$V_{clu-I}(\hat{t}) = (8 \times 4)^2 (1 - 2/8) S_b^2 / 2 = 32^2 \times 3/4 \times 340^2 / 2 = 6\ 663^2。$$

群间方差 $S_b^2 = 340^2$ 也将用于二级整群抽样中。因此，总和估计值 $\hat{t}$ 的设计效应为：

$$\text{DEEF}_{clu-I}(\hat{t}) = \frac{V_{clu-I}(\hat{t})}{V_{srs}(\hat{t})} = \frac{6\ 663^2}{7\ 283^2} = 0.84。$$

在这个例子中，一级整群抽样比 SRSWOR 设计略微有效。但是，在复杂调查中，使用估算的设计效应来评估时，由于有群内相关，整群抽样通常比 SRSWOR 低效。这在后面的章节中可以看到。这一预期外的结果可以部分地为从行政上划分整群所解释，这样的方法生成了在 UE91 差异上内部相对同质的结果。如果使用其他标准来划分整群，比如上班路程的远近，可能会得到不同的结果。这是因为，与区域相邻的自治市相比，这种情形下的失业情况同质性更高。

下面使用 1991 年省级人口数据的一级整群抽样的例子中，根据估算方差来划分整群，得到的结果比 SRSWOR 更为低效。但是，这样的结果在很大程度上取决于样本的组成。因为，从规模较小、异质性较高的总体中只抽取了两个整群。

范例 3.6

1991 年省级人口数据一级整群样本的分析。1991 年省级人口数据按区域邻近被分成 8 个（$M=8$）整群，每一个含有 4 个（$B=4$）相邻的自治市。样本需要 8 个自治市，所以元素样本为 8。由于整群规模相等，所以整群样本为 $m=2$。使用无放回式简单随机抽样抽取整群的样本。我们抽取了第 2 和第 8 整群。表 3.7 给出了这两个整群中的 8 个自治市。

表 3.7　1991 年省级人口的一个一级 CLU 样本（含两个整群；抽中整群用黑体）

整群标识			抽中整群中 UE91 的均值与总和	
STR	CLU	元素（含市）	均值 $\bar{Y}_i$	总和 T_i
1	1	Jyväskylä, Korpilahti, Muurame, Säynätsalo	…	…
1	**2**	**Jämsä, Jämsänkoski, Keuruu, Kuhmoinen**	**535.25**	**2 141**
1	3	Saarijärvi, Konginkangas, Äänekoski, Sumiainen	…	…
1	4	Kannonkoski, Karstula, Kyyjärvi, Pylkönmäki	…	…
1	5	Suolahti, Hankasalmi, Konnevesi, Laukaa	…	…
1	6	Joutsa, Leivonmäki, Luhanka, Toivakka	…	…
1	7	Jyväskylä mlk., Multia, Petäjävesi, Uruainen	…	…
1	**8**	**Kinnula, Kivijärvi, Pihtipudas, Viitasaari**	**289.00**	**1 156**

抽样比例（整群）：$m/M = 2/8 = 0.25$；

“…”未抽中整群。

分析数据所需的样本标签包括以下 3 个变量:STR 是层级标签,由于整群总体没有分层,在这里是常数,即这里只有一个层级;整群标签(2 或 8)用变量 CLU 来表示;而权重变量是一个常数 WEIGHT = 4,亦即整群规模。整群层次上有限总体的校正比例为 1 - 0.25 = 0.75,因而抽样比例为 0.25。

表 3.8 1991 年省级人口数据一个一级 CLU 样本(n = 8)的估计值

统计量	变 量	参 数	估计值	s. e	c. v	deff
总和	UE91	15 098	13 188	3 412	0.26	1.92
比率	UE91, LAB91	12.65%	12.93%	0.6%	0.04	1.44
中位值	UE91	229	337	132	0.39	1.29

表 3.8 给出了估算结果总和 $\hat{t}$、比率 $\hat{r}$ 与中位值 $\hat{m}$ 以及它们所对应的总体参数 T,R 和 M。这里可以看出,对于 3 个估计值而言,一级整群抽样都显得低效。deff 的估计值大于 1(1.29≤deff≤1.92)。另外,估算的 deff($\hat{t}$) = 1.92,与相应的参数 DEFF($\hat{t}$) = 0.84 相去甚远。这是由于整群数目太少,因而造成了估算设计方差时的不稳定。方差估计值在很大程度上取决于哪些整群被抽中,因而,如果是另外两个整群被抽中,有可能得到较小的 deff。第 5 章将讨论这一不稳定的问题。

两级整群抽样

面对规模较大的整群时,在抽中整群中做次一级抽样也是常见的。比如,当整群规模不同时,这样的方案提供了控制元素层次样本规模 n。同时,在给定样本规模的情形下,与一级整群抽样相比,由于使用再次抽样,可以提高抽取的整群的数目,因而提高效率。实际的动机是,仅仅在抽中的整群中需要抽样框。

在两级整群抽样的第一级抽样中,使用标准的元素层次的抽样技术,例如 SRSWOR,SYS 或是 PPS,从总体整群中抽取一个整群样本——主要抽样单位(PSUs)。同时,也可以将总体中的整群根据现成的辅助信息分层。实际中,经常使用的是最简单的两级分层整群抽样设计。其中,在各个层级中抽取两个整群。这样可以使用更多数目的层级,以提高效率。在第二级抽样中,继续使用标准的元素层次的抽样技术,在抽中整群中抽取元素。在实际中,总体中整群的规模以及整群样本的大小通常并不相同。同时,在抽样的各个阶段,选中概率可以不同。但是,通过小心选择在各个阶段的抽样比例与抽样技术,可以得到一个整体抽样比例为常数的样本。这样的多级抽样设计被称为等概设计。

在一级抽样中,抽中整群中所有元素都纳入到元素层级的样本。因此,仅仅由抽样产生的离异是群间差异。但在两级整群抽样中,增加了另一个离异

的来源——再次抽样，即是群内差异。这也增加了总体上的离异。

为了演示两级整群抽样的基本情况，我们假定在两级抽样中均使用SRSWOR方法，以及总体中 M 个整群规模相同，即对于所有的 $i, B_i = B$。元素层次的总体规模是 $N = M \times B$。同时，为了简便，让我们进一步假定所有 m 个样本整群的规模相同，即，$b_i = b$。因而，元素层次的样本规模为 $n = m \times b$。这些假设之下的整群抽样将得到，总体整群与总体元素均相等的选中概率，即是等概样本。可以看出，一级抽样的抽样比例为 m/M，二级抽样的为 b/B，而整体上的抽样比例为 $(m/M) \times (b/B) = n/N$。

估算的主要目的通常在第二级，即元素层次的参数。让我们讨论总体合计 $T = \sum_{i=1}^{M} T_i$。其中，$T_i = B \times \overline{Y}_i$ 是整群 i 中的总体合计；$\overline{Y}_i = \sum_{i=1}^{B} \overline{Y}_{ik}/B$ 是整群 i 中的元素均值。总和 T 的无偏估计量是：

$$\hat{t} = (M \times B) \sum_{i=1}^{m} \bar{y}_i/m, \tag{3.13}$$

其中，$\bar{y}_i = \sum_{k=1}^{b} y_{ik}/b_i$ 是样本整群 i 中元素均值。在推导 $\hat{t}$ 的设计方差中，可以将总体方差分解成组间和组内方差。估计值 $\hat{t}$ 的设计方差是群间方差 S_b^2 与群内方差 S_w^2 的加权和：

$$V_{clu-II}(\hat{t}) = (M \times B)^2 \left[\left(1 - \frac{m}{M}\right) \frac{S_b^2}{m} + \left(1 - \frac{b}{B}\right) \frac{S_w^2}{mb} \right], \tag{3.14}$$

与

$$S_b^2 = \frac{1}{M-1} \sum_{i=1}^{M} (\overline{Y}_i - \overline{Y})^2,$$

$$S_w^2 = \frac{1}{M(B-1)} \sum_{i=1}^{M} \sum_{k=1}^{B} (Y_{ik} - \overline{Y}_i)^2,$$

$\overline{Y} = T/(M \times B)$ 是每一元素的总体均值。群间方差是由第一级整群抽样产生的，它与一级整群抽样相似。额外的群内方差是由再抽样产生的。在一级整群抽样中，群内方差部分为0。这是因为抽中整群中的所有 B 个元素均纳入了样本中，即 $b = B$。

方差 S_b^2 与 S_w^2 的估计值通过用样本取值置换总体取值获得。我们有，

$$\hat{s}_b^2 = \frac{1}{m-1} \sum_{i=1}^{m} (\bar{y}_i - \bar{y})^2,$$

$$\hat{s}_w^2 = \frac{1}{m(b-1)} \sum_{i=1}^{m} \sum_{k=1}^{b} (y_{ik} - \bar{y}_i)^2,$$

其中，$\bar{y} = \sum_{i=1}^{m} \bar{y}_i/m$ 是样本的元素均值。$\hat{t}$ 的设计方差的估算公式则有，

$$\hat{v}_{clu-II}(\hat{t}) = (M \times B)^2 \left[\left(1 - \frac{m}{M}\right) \frac{\hat{s}_b^2}{m} + \left(1 - \frac{b}{B}\right) \frac{m}{M} \frac{\hat{s}_w^2}{mb} \right]。 \tag{3.15}$$

从式 3.15 中可以推出，当第一级抽样中的比例 m/M 较小时，方差估计值的第二部分可以忽略不计。那么，只含有群间离异的方差可以当成 $\hat{t}$ 的设计方差的略微负向偏离的近似值。它可以很容易地由整群层次的数值计算出来。同时，当 m/M 较小时，第一级的有限总体校正接近于 1，因而可以忽略，得出了一个放回式方差估计值。在后面章节讨论调查分析中，这样的方差近似法将被广泛地使用。另一方面，当 m/M 较大时，群内方差可能是方差估计值较大的组成部分。

在实际中，组群规模 B_i 与抽中整群中的样本规模 b_i 均不相同，而整群总体也可以划分层级。考虑到分层和整群样本的差异，需要恰当地使用总和以及总和估计值的设计方差的估算量。对于总和而言，可以使用比率类型的估计量或是基于 PPS 整群抽样的估计量。其中，整群规模被用作辅助规模变量。第 5 章将讨论，两级分层整群抽样情形下的比率类型估计量的设计方差的估算。那时，将介绍各种近似的方差估计量。

使用 PPS 抽取整群可以控制麻烦的整群规模的不同。让我们假设，需要一个固定规模 n 的等概样本。通过使用选中概率与整群规模 B_i 成比例的 PPS 抽取 m 个整群，再从抽中整群中抽取等数目的元素 $b_i = b$，可以得到这样一个样本。选中概率可以从下面的公式中推出：

$$\frac{n}{N} = \frac{m \times B_i}{\sum_{i=1}^{M} B_i} \times \frac{b}{B_i},$$

其中，m 是所需的样本整群的数目，$b = n/N \times \sum_{i=1}^{m} B_i/m$。

在下一个例子中，通过计算 DEFF，我们来评估等规模整群的简单情形下的两级 CLU 设计的效率。

范例 3.7

1991 年省级人口数据两级整群抽样的效率。含有相邻自治市的整群的数目为 8，因而 $M=8$。每个整群中有 $B=4$ 个自治市。我们比较在估算总和 T 时，一级与两级整群抽样的效率。两种设计在第一级抽样中面对的都是等规模整群。在一级整群抽样中，抽取 2 个整群（$m=2$），其中所有 4 个自治市均纳入元素层次的样本中。样本规模为 $n = m \times B = 2 \times 4 = 8$。在两级整群抽样中，第一级样本抽取 $m=4$ 个整群；在第二级中，从这 4 个整群中各抽取 $b=2$ 个自治市。元素层次的样本规模也是 $m \times b = 4 \times 2 = 8$。

在一级 CLU 设计中，设计方差为 $V_{clu}(\hat{t}) = 6\ 663^2$，设计效应 $\mathrm{DEFF}(\hat{t}) = 0.84$。在两级 CLU 设计中，我们必须首先计算群间与群内方差。范例 3.3 中计算出的群间方差为 $S_b{}^2 = 340^2$。群内方差为：

$$S_w^2 = \frac{1}{8(4-1)} \sum_{i=1}^{8} \sum_{k=1}^{4} (Y_{ik} - \bar{Y}_i)^2 = 660^2。$$

$\hat{t}$ 的设计方差则是：

$$V_{clu-II}(\hat{t}) = (8\times4)^2\left[\left(1-\frac{4}{8}\right)\frac{340^2}{4}+\left(1-\frac{2}{4}\right)\frac{660^2}{4\times2}\right] = 6\ 532^2$$

两级整群抽样中的 $\hat{t}$ 的 DEFF 是：

$$\mathrm{DEFF}_{clu-II}(\hat{t}) = 6\ 532^2/7\ 283^2 = 0.80。$$

与一级 CLU 设计相比，两级设计略微更有效率。部分原因是由于两级设计的特征所决定的。给定 n，可以抽取比一级设计更多的第一级单位（整群——译者注）。在这个例子中，整群样本的数目加倍，因而降低了第一级方差。在总体方差中，归因于第一级（群间）的占 35%，归因于第二级（群内）的占 65%。因而，群内部分占主要份额。部分原因是整群内异质性相对较高。但是，应当注意到，在 1991 年省级人口数据中，整群数目较小，而整群样本规模及再抽样的规模也较小。所以，这些计算仅仅是假想的例子。在实际调查中，这些相应的数字要大得多，整群也相对同质，设计方差的主要份额则由群间差异造成。

下一个例子展示从 1991 年省级人口数据中，使用两级 CLU 设计抽取样本的计算结果。通过估算的设计方差，我们将讨论效率。也将同范例 3.6 中的一级 CLU 设计比较效率。

范例 3.8

1991 年省级人口数据两级 CLU 样本分析。在第一级中，编号为 2，3，4 和 7 的整群被抽中。在第二级中，每一抽中整群中有两个自治市被抽中。表 3.9 给出了整群总体与两级 CLU 样本。

表 3.9　1991 年省级人口的一个两级整群样本

整群标识			抽中整群中 UE91 的均值与总和	
STR	CLU	元素（含市）	均值 $\bar{y}_i$	总和 $\hat{t}_i$
1	1	Jyväskylä，Korpilahti，Muurame，Säynätsalo	…	…
1	2	Jämsä，Jämsänkoski，**Keuruu**，**Kuhmoinen**	473.5	1 894
1	3	Saarijärvi，**Konginkangas**，**Äänekoski**，Sumiainen	454.5	1 818
1	4	Kannonkoski，Karstula，**Kyyjärvi**，**Pylkönmäki**	96.0	384
1	5	Suolahti，Hankasalmi，Konnevesi，Laukaa	…	…
1	6	Joutsa，Leivonmäki，Luhanka，Toivakka	…	…
1	7	Jyväskylä mlk.，Multia，**Petäjävesi**，**Uruainen**	241.0	962
1	8	Kinnula，Kivijärvi，Pihtipudas，Viitasaari	…	…

第一级：4 个整群（2，3，4 与 7）的 SRSWOR 样本；

第二级：抽中的 4 个整群中两个元素；

抽样比例：第一级 4/8 = 0.50，第二级 2/4 = 0.50；

“…”未抽中整群；

抽中的元素用黑体。

在分析两级 CLU 设计数据时,需要以下设计标签:层级标签 STR——对于所有样本元素均为常数 1;整群标签 CLU——对应抽中整群的编号 2,3,4,7;权重变量 WEIGHT——对于所有选中元素均为常数 4。应当注意到,当整群规模不同以及各整群的选择比例不等时,权重也会有不同。由于在两级抽样中使用了 SRSWOR,我们有第一级抽样比例 4/8,第二级抽样比例 2/4,得出所有样本元素的权重是 $w_{ik}=(M\times B)/(m\times b)=(8\times4)/(4\times2)=4$。表3.10给出了失业人数 $\hat{t}$、失业比率 $\hat{r}$ 与失业中位值 $\hat{m}$,以及相应的总体参数 T,R 和 M。

表 3.10　1991 年省级人口中一个两级 CLU 样本($n=8$)的估计值

统计量	变　量	参　数	估计值	s. e	c. v	deff
总和	UE91	15 098	10 116	2 659	0. 26	0. 93
比率	UE91, LAB91	12. 65%	13. 81%	0. 5%	0. 04	0. 99
中位值	UE91	229	192	49	0. 25	0. 84

总和、比率与中位值估计值的估算设计效应(deff)接近 1,显示两级 CLU 设计的效率与 SRSWOR 相去不远。但是,与所有估计量的设计效应估计值远大于 1 的一级 CLU 设计相比,效率差别较大。在一级设计中,整群样本较小,因而造成方差估计时的严重不稳定性。另一方面,在两级设计中,选中了整群总体的一半。因而,设计并非那样地不稳定。另外,总体整群的异质性也较高。应当注意到,这个例子中的整群划分仅仅是为了演示两级整群抽样,而并非实际调查中的整群抽样。在后面的章节中,将讨论这样的实际中的例子。

群内相关与效率

整群抽样的效率对于整群内部的组成相当敏感。当整群内部异质性较高而每一整群能够代表总体的整体组成时,整群抽样与简单随机抽样同等高效。当整群内部同质,并且群间差异较大时,效率会降低。在实际中,许多自然形成的总体次级整群是后一种情形。

可以通过测量整群同质性的群内相关来研究效率。这一相关系数可以纳入整群抽样估计量的设计方差的方程中。回忆一下系统抽样中,相似的系数(组内相关系数)也扮演了重要角色;SYS 可以被看成是只抽取了一个整群的一级整群抽样的特例。

让我们假定总体整群的规模相同 $B_i=B$。我们首先讨论 ANOVA 并分解研究变量 y 的离异总和 SST 为群内离异(SSW)与群间离异(SSB)。总和离异 SST 可以写成,

$$\sum_{i=1}^{M}\sum_{k=1}^{B}(Y_{ik}-\overline{Y})^2=\sum_{i=1}^{M}\sum_{k=1}^{B}(Y_{ik}-\overline{Y}_i)^2+\sum_{i=1}^{M}B(\overline{Y}_i-\overline{Y})^2,\qquad(3.16)$$

其中,与前面一样,Y_{ik}是研究变量在整群 i 中元素 ik 的总体取值,$\overline{Y}$ 是元素的

整体均值，$\bar{Y}_i$ 是元素的整群均值。

在整群抽样中，使用章节 2.4 中 SYS 情形下推导出的组内相关系数 ρ_{int}，我们有，

$$\rho_{int} = 1 - \frac{B}{B-1} \times \frac{SSW}{SST}。 \tag{3.17}$$

群内相关系数的解释取决于两个方差部分在整体离异中的份额。首先，当所有离异来源于整群内部而非整群外部时，群内相关系数最小 $\rho_{int} = -1/(B-1)$；另一方面，当所有离异来源于整群外部时，即整群内部完全同质，相关系数最大 $\rho_{int} = 1$。在 $\rho_{int} = 0$ 的情形下，元素随机地散布于各整群中。

让我们讨论与 SRSWOR 相比的、样本规模同为 n 的一级整群抽样的效率。在 CLU 设计中的总和 T 的估计值 $\hat{t}$ 的设计方差在式 3.12 中给出，

$$V_{clu-I}(\hat{t}) = (M \times B)^2\left(1 - \frac{m}{M}\right)\frac{S_b^2}{m},$$

其中，S_b^2 是群间方差部分。从等式 3.16 和 3.17 中群间方差部分可以写成

$$SSB = \frac{SST}{B}[1 + (B-1)\rho_{int}]。$$

将它带入上面的方差公式中，我们有

$$V_{clu-I}(\hat{t}) = (M \times B)^2\left(1 - \frac{m}{M}\right)\frac{S^2}{m}\left[\frac{1}{B}(B-1)\rho_{int}\right] \times \frac{N-1}{N} \times \frac{M}{M-1}。$$

假定 N 和 M 数目巨大，最后两项接近 1，可以不计。因此，由于 $m \times B = n$，我们得到一个基于总体方差与群内相关系数 ρ_{int} 的 $\hat{t}$ 的设计方差：

$$V_{clu-I}(\hat{t}) \approx (M \times B)^2\left(1 - \frac{m}{M}\right)\frac{S^2}{n}[1 + (B-1)\rho_{int}], \tag{3.18}$$

但是，与此相应的 $\hat{t}$ 的 SRS 设计方差可以写成：

$$V_{srs}(\hat{t}) = (M \times B)^2\left(1 - \frac{n}{N}\right)\frac{S^2}{n},$$

由上得到 $\hat{t}$ 的 DEFF 为：

$$\text{DEFF}_{clu-I}(\hat{t}) = \frac{V_{clu-I}(\hat{t})}{V_{srs}(\hat{t})} = 1 + (B-1)\rho_{int}, \tag{3.19}$$

其中，在有限总体校正项 $V_{clu-I}(\hat{t})$ 中，$m/M = n/N$。

DEFF 的等式显示，当 ρ_{int} 为正时，通常实际情况也是如此，整群抽样比简单随机抽样低效。给定 ρ_{int}，DEFF 随着整群数目 B 的增加而增加。在本小节的最后的例子中，我们将进一步讨论作为整群规模与群内相关系数的函数的效率。

范例 3.9

1991 年省级人口数据的群内相关系数、整群规模以及 DEFF。我们使用 8 个区域整群作为类别因子，计算变量 UE91 的一元方差分析。表 3.11 给出了

结果。

表 3.11 1991 年省级人口数据中一级整群抽样($M=8,B=4$)的人口 ANOVA 表

离异来源	df	平方和	平均平方和
群间	7	$SSB=32.30\times10^5$	$MSB=4.61\times10^5$
群内	24	$SSW=139.02\times10^5$	$MSW=5.79\times10^5$
合计	31	$SST=171.32\times10^5$	$S^2=5.53\times10^5=743^2$

将数据代入等式 3.17 中,我们有:

$$\rho_{int}=1-\frac{4}{4-1}\times\frac{139.02\times10^5}{171.32\times10^5}=-0.082。$$

设计效应可以根据公式 3.19 近似得出(假定整群数目相同 $B=4$),$DEFF_{clu-1}=1+(B-1)\rho_{int}=1+(4-1)(-0.082)=0.754$。

这一数字小于范例 3.4 中计算出的真实值 DEFF=0.84。这是因为公式 3.17 是一个近似算法,更适合于抽样比例较小的大规模总体。

看起来,组内或是群内相关系数在系统抽样和整群抽样中是一个重要的设计参数。组内相关系数测量总体内次级组别内成对元素的相关性。在 SYS 中,次级组别是抽样间隔内的元素。在整群抽样中,群内相关系数表示同一整群或是自然形成的次级总体内元素的相互依赖程度。我们将讨论若干这样的分组结构:学校里的学生、公司里的员工以及家庭户里的家庭成员。使用群内相关系数来测量内在同质性时,有几个可选方案。在系统抽样与整群抽样中,使用基于设计的方法来计算群内相关系数。在多变量模型中,其他方案更合适。这包括在第 8 章多变量调查分析估算中引入的“有效”的群内相关系数。在章节 9.4 中,使用基于模型的方法来计算组内相关系数。这也适用于另一种整群构成的方案——根据访谈员工作量来划分(参见章节 9.1)。

小 结

由于许多总体本身已有自然划分的次级整群,整群抽样在实际中经常使用。实际中典型的整群包括行政区域单位、城市街区或街区类单位、家庭户、公司以及学校或学校内的年级。由于实际和经济的原因,这样的整群在抽样和收集数据中经常会使用到。实际的原因是,只需要准备再抽样中抽中整群的抽样框。经济的原因是,整群抽样的费用效率很高。本书后面有各种整群抽样设计的例子。整群抽样的缺点是,在实际中整群内部的相对同质,因而造成统计效率不及简单随机抽样。但是,较高的费用效率可以成功弥补这样的不足。

作为我们的演示数据,1991 年省级人口数据看起来有局限,并不能完全演示整群抽样,只用来演示一级和两级整群抽样和估算过程中的基本原则。

在大规模调查中，通常在总体和样本中有数目较大的整群。同时，整群总体也可以被分层，抽样可以通过几个步骤完成。在分析这样的数据时，估算过程中通常使用有着近似方差的比率类估计量。第 5 章将详细讨论这些议题。

绝大多数调查抽样的教科书均讨论了整群抽样。更多的参考书包括，基什(Kish,1965)、洛尔(Lohr,1999)、列维与莱密肖(Levy and Lemeshow,1991)，以及斯利基德斯与博斯卡(Snijders and Bosker,2002)。这些书涵盖了整群抽样中解释性的、高级的以及更加理论性的议题。

3.3 模型辅助估算

引　言

迄今讨论的技术中，总体元素的辅助信息被用在抽样阶段，以获取一个高效的抽样设计。我们现在转向另一个使用辅助信息的方式。我们的目标是介绍可以获取感兴趣的总体参数的更好的估计值的估算方法。这是相对于基于抽样设计的估算方法计算出的估计值而言的。

让我们假定总体中已有一组辅助变量所表示的合适的辅助数据。在这些变量中，一些是定类变量，另一些是定序变量。一些辅助数据用于抽样过程中，另一些则可以用来提高效率，比如，使用的方式是，使用与研究变量 y 相关的辅助变量 z 来降低原有的 y 的总体总和的设计方差。在桑德尔等(Sarndal et al. 1992)中，基于设计的模型辅助估算中讲解了这样的技术。模型辅助估算是指使用线性回归等模型的估算公式的一种特征。这一特征在估算有限总体中如总和等感兴趣的参数时，纳入了辅助信息。模型辅助估算应当有别于第 8 章讨论的多变量调查分析方法。在那里，使用模型的目的是多变量调查分析。

下面给出一个简要的模型辅助估算的回顾。具体地讲，我们将讨论后续分层、比率估算与回归估算。这些方法是所谓的通用回归估算的特例。所有这些方法都通过使用总体中已有的辅助信息来提高估算效率。这样可以使得估计值接近于总体中的真值，并降低从样本数据中计算估计值的设计方差。

在模型辅助估算中，需要一个与研究变量 y 相关的辅助变量 z。当这一变量是类别变量时，可以将目标总体 U 根据一定的原则，分割成 $U_1,\cdots,U_g,\cdots,U_G$。在后续分层中，这些次级总体被称作后续层级。如果这些后续层级同质性较高，则这样的分割能够抓住研究变量 y 的较大份额的方差，降低估计值的基于设计的方差。同时，后续分层也可以用来提高点估计的精度，以及降低由无回答导致的样本估计偏差。

辅助变量 z 通常是连续性的。当它与研究变量 y 强相关时，可以设定一个以 y 为因变量、z 为预测变量（亦即自变量——译者注）的线性回归模型。这一模型可以用观测到的样本来估算，并用于原有总体参数的估计中。比率与回归估算可以用于此处。使用这些方法，通常可以提高效率和增加精确度。

为了建立一个模型辅助的估算公式，可以考虑两种权重。初级权重通常是抽样权重 w_k，它就是选中概率 π_k 的倒数。本书广泛使用这样的权重。另一种权重被称作 g-权重，它的取值 g_k 取决于选中样本以及选中的估算公式。乘积 $w_k^* = g_k w_k$ 产生一个新的权重，被称作标准化权重。它通常用于模型辅助的估算过程中。使用标准化权重，模型辅助公式可以写成 $\hat{t}_{cal} = \sum_{k=1}^{n} w_k^* y_k$。标准化权重有一个特征，对于比率估算而言，辅助变量 z 的总和估计值 $\hat{t}_{z,cal} = \sum_{k=1}^{n} w_k^* z_k$ 与已知总体总和 T_z 完全一致。在后续分层、比率估算与回归估算中，g-权重与标准化权重将被明确提出来。

虽然，模型辅助估算在实际中通常是在更为复杂的设计中的使用，但我们将在 SRSWOR 设计中方便地介绍它的基本原则。另一个简化是，只使用一个辅助变量。同时，正如回归估算的讨论中，如果存在多个辅助变量，这一假设可以变通。在同时选择抽样设计与合适的估算公式时，我们将使用估算策略的概念。表 3.12 中列出了将要讨论的模型辅助策略。基于设计的策略中，并不需要辅助信息。

表 3.12　总体总和的估算策略

策　略	辅助信息		辅助模型
	基于设计的策略		
SRSWOR	未用		无
SRSWR	未用		无
	模型辅助策略		
后续分层	SRS* pos	间断	ANOVA
比率估算	SRS* rat	连续	回归（无截距）
回归估算	SRS* reg	连续	回归

后续分层

当已有非连续的辅助变量时，后续分层可以用来提高估算效率。在样本抽取之后，这个变量用来将样本分层。回忆在章节 3.1 中，作为抽样设计中的一部分，分层通常会提高效率。这是通过谨慎选择分层变量，使得研究变量 y 层级内的差异较小而获得的。后续分层有着同样的目的。为了与前期分层相区隔，总体分成 G 个后续层级。

为了实施后续分层，首先，将样本数据与从行政登记或是官方统计中得来的辅助信息合成一体。在将样本数据与后续层级信息以及相应的选中概率合在一起后，我们可以像在普通前期分层中一样，使用同样的方法来进行估算。当然，也有一定的差别。由于我们在样本抽取后，或是更经常地，在数据收集之后分层，我们不能设定任何配额方案。样本规模 n 是设定的。但是，它的配额方案在样本抽取前是未知的。这样的特征在估算总和时并不会造成问题。但在估算总和的方差时，则需要更多的关注。

y 的总和 T 的后续估算公式如下：

$$\hat{t}_{pos} = \sum_{g=1}^{G} \hat{t}_g = \sum_{g=1}^{G} \sum_{k=1}^{n_g} w_{gk}^* y_{gk}, \tag{3.20}$$

其中，$\hat{t}_g = N_g \bar{y}_g$ 是后续层级总和 T_g 的估计值，N_g 是后续层级 g 的规模。后续层级权重为 $w_{gk}^* = g_{gk} w_{gk}$，其中的 g-权重为 $g_{gk} = N_g / \hat{N}_g$，分母是估算的后续层级的规模。w_{gk}是原有的抽样权重。w_{gk}^*的计算将在例子 3.9 中列出。根据对于获取样本的构成的使用方法，有不同的方法来确定 $\hat{t}_{pos}$的方差。样本的构成是指，后续层级的样本规模 n_g 的实际分布。如果这一构成是给定的，那么条件方差与通常的分层样本中的一样：

$$V_{srs,con}(\hat{t}_{pos} \mid n_1, \cdots, n_g, \cdots, n_G) = \sum_{g=1}^{G} N_g^2 \left(1 - \frac{n_g}{N_g}\right) \frac{S_g^2}{n_g}, \tag{3.21}$$

其中，后续层级的方差为 $S_g^2 = \sum_{k=1}^{N_g} (Y_{gk} - \bar{Y}_g)^2 / (N_g - 1)$ 。对于所有可能的 n 的构成组合（即，样本数目 n 在各个层级中可能的分布——译者注），取式 3.21 的平均值，就可以得到非条件方差。这样的另一个方差公式为：

$$V_{srs,unc}(\hat{t}_{pos}) = \sum_{g=1}^{G} N_g^2 \left(1 - \frac{E(n_g)}{N_g}\right) \frac{S_g^2}{E(n_g)}, \tag{3.22}$$

其中，$E(n_g)$是层级规模的期望值。这一方差有多种近似值。其中一个为：

$$V_{srs,unc}(\hat{t}_{pos}) \approx N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \left[\sum_{g=1}^{G} \left(\frac{N_g}{N}\right) S_g^2 + \frac{1}{n} \sum_{g=1}^{G} \left(1 - \frac{N_g}{N}\right) S_g^2 \right]。 \tag{3.23}$$

在样本规模较小时，条件方差与非条件方差的差别可能相当大。用 $\hat{s}_g^2$ 替代 S_g^2，可以得出相应的方差估计值 $\hat{v}_{srs.con}(\hat{t}_{pos})$ 与 $\hat{v}_{srs.unc}(\hat{t}_{pos})$。而 $\hat{s}_g^2 = \sum_{k=1}^{n_g} (y_{gk} - \bar{y}_g)^2 / (n_g - 1)$ 。为了演示的目的，在下一个例子中，我们将计算 $V_{srs.con}$ 与 $V_{srs.unc}$。

范例 3.10

后续分层估算。这里使用的样本是章节 2.3 中用 SRSWOR 从 1991 年省级人口数据中抽取的（见范例 2.1）。根据自治市的行政区划，将样本后续分层为城镇与农村。目标总体中有 7 个城镇 $N_1 = 7$，25 个农村 $N_2 = 25$。后续层

级的取值,1 为城镇,2 为农村。

表 3.13 给出了后续分层的估算观察中使用的样本信息。

让我们进一步考虑总和 T 的估算。从表中,UE91 的后续层级总和的估算是 $\hat{t}_1 = N_1\bar{y}_1 = 7 \times 1\ 868 = 13\ 076$,以及 $\hat{t}_2 = N_2\bar{y}_2 = 25 \times 201.2 = 5\ 030$。使用这些估计值,总和 T 的后续分层估计值为 $\hat{t}_{pos} = \hat{t}_1 + \hat{t}_2 = 18\ 106$。

另外的,总和估计值 $\hat{t}_{pos}$ 也可以通过后续层级的权重 w_k^* 计算出来。为了计算 w_k^*,应当使用随样本变化的 g_k 权重来校正原来的抽样权重 w_k。为此,要先决定后续层级规模的估计值。后续层级 g 中的样本元素的元素权重表示为 w_{gk},后续层级规模的估计值 $\hat{N}_g$ 就是这些权重的和。因此,后续层级 g 中元素 k 的 g 权重即是 $g_{gk} = N_g/\hat{N}_g$,其中,N_g 是后续层级 g 的准确规模。比如,表 3.13中,SRS 情形下的原始抽样权重是 $w_k = 4$,这对每一总体元素是相同的。在第一个后续层级中,其规模 $N_1 = 7$,其估计的规模为 $N_1 = 4 + 4 + 4 = 12$。这是因为,这一后续层级中有 3 个抽中的元素。那么,相应的 g 权重为,$g_{1k} = N_1/\hat{N}_1 = 7/12 = 0.583\ 3$。最后,第一后续层级的层级权重为 $w_{1k}{}^* = g_{1k} \times w_{1k} = 0.583\ 3 \times 4 = 2.333\ 3$。这一数值对于第一后续层级中的所有元素都是一样的(城镇自治市)。使用后续层级权重,$\hat{t}_{pos}$ 的估计值与前面计算的相同。

表 3.13　1991 年省级人口中后续分层加权的 SRSWOR 样本

样本设计标识			元素标签	研究变量		后续分层		
							g	Post.
STR	CLU	WGHT		UE91	LAB91	POSTSTR	WGHT	WGHT
1	1	4	Jyväskylä	4 123	33 786	1	0.583 3	2.333 3
1	4	4	Keuruu	760	5 919	1	0.583 3	2.333 3
1	5	4	Saarijärvi	721	4 930	1	0.583 3	2.333 3
1	15	4	Konginkangas	142	675	2	1.250 0	5.000 0
1	18	4	Kuhmoinen	187	1 448	2	1.250 0	5.000 0
1	26	4	Pihtipudas	331	2 543	2	1.250 0	5.000 0
1	30	4	Toivakka	127	1 084	2	1.250 0	5.000 0
1	31	4	Uurainen	219	1 330	2	1.250 0	5.000 0

计算无条件方差抽样比例:8/32 = 0.25。

计算条件方差抽样比例:

层级 1(城市) = 3/7 = 0.43;

层级 2(农村) = 5/25 = 0.20。

表 3.14 给出了总和与比率的估算结果。其中保留了原有的样本标签,如 STR = 1 与 CLU = ID。但是,元素权重由后续层级权重所代替,第一个后续层级的抽样比例为 0.43,第二个后续层级的抽样比例为 0.20。同时,使用了原

有的抽样权重。在两个后续层级中,估算无条件方差的抽样比例均为 0.25。注意,这一程序与公式 3.23 大体近似。为了便于比较,表中也列出了在 SRSWOR 情形下,基于设计的估计值 $\hat{t}$ 与 $\hat{r}$。

表 3.14 1991 年省级人口中一个 SRSWOR 样本的后续分层估计值

(1)后续分层估计值(条件)					
统计量	变　量	估计值	s.e	c.v	deff
总和	UE91	18 106	6 014	0.33	0.33
比率	UE91,LAB91	12.97%	0.45%	0.03	0.59
(2)后续分层估计值(无条件)					
统计量	变　量	估计值	s.e	c.v	deff
总和	UE91	18 106	7 364	0.41	0.50
比率	UE91,LAB91	12.97%	0.49%	0.03	0.70
(3)基于设计的估计值					
统计量	变　量	估计值	s.e	c.v	deff
总和	UE91	26 440	13 282	0.50	1.00
比率	UE91,LAB91	12.78%	0.41%	0.03	1.00

这一比较表明了后续分层对点估计的影响。在估算总体总和时获利颇多。失业人数的估计值是 $\hat{t}_{pos}=18\ 106$,比基于设计的估计值 $\hat{t}=26\ 440$ 更接近于真实值 $T=15\ 098$。而比率的估计值仅有很小的变化。

总和估计值更加准确的原因很显然。在 SRSWOR 情形下,从城镇与农村中抽取的样本与它们各自的比例相当,城镇 $(8/32)\times 7\approx 2$,农村 $(8/32)\times 25\approx 6$。城镇自治市人口更多,失业人数也更多。如果碰巧它们在样本中占较大的份额,则基于设计的估计值将夸大人口总和。而后续分层可以校正(至少是部分地)这一偏差。所以,我们可以得到一个更接近于真实值的点估计。

后续分层也可以提高效率。对于总和而言,更是如此。与单纯基于设计的估计值 $\hat{t}$ 相比,$\hat{t}_{pos}$ 的条件方差的估计值只有三分之一,由 deff = 0.33 来表示。如果是非条件情形下,则 deff = 0.50。非条件方差估计值大于条件方差估计值。这是因为,由定义,后续层级的样本规模 n_g 是随机变量,它们的方差增加了总的方差。

总体总和的比率估算

在前面的后续分层的情形下,使用样本数据与一个辅助变量讨论了研究变量 y 的总体总和 T 的估算。当已有连续性辅助变量 z 时,比率估算也可以提高效率。这样的方法需要人口总和 T_z 以及 z 的 n 个样本取值 z_k。这样的信息,通常可以从行政登记或是官方统计中得到。它用来提高对 T 的估算。首

先，计算出比率 $R = T/T_z$ 的样本估计值 $\hat{r} = \hat{t}/\hat{t}_z$。然后，用 $\hat{r}$ 乘以已知的总和 T_z。当研究变量与辅助变量的比率 Y_k/Z_k 在总体中接近于常数时，总和的比率估算非常有效。

比率估计值通常十分有效，但略含偏差。由于有偏差，所以在考察抽样误差时，使用平均差方(MSE)，而不使用方差。已经给出比率估计值的偏差为 $1/N$，所以增加样本规模降低偏差。因而，在大样本中，方差可以成为 MSE 的近似值。经典抽样理论广泛地研究了比率估计的特征。

让我们考虑，在无放回式简单随机抽样情形下的 y 的总和 T 的比率估算。我们对于以下的比率估算总和感兴趣：

$$\hat{t}_{rat} = \hat{r} \times T_z = \sum_{k=1}^{n} w_k^* y_k, \tag{3.24}$$

其中 $\hat{r} = \hat{t}/\hat{t}_z = N\bar{y}/N\bar{z} = \sum_{k=1}^{n} y_k / \sum_{k=1}^{n} z_k$，$T_z$ 是辅助变量 z 的总体总和。标准权重 $w_k^* = g_k w_k = (T_z/\hat{t}_z) w_k$。

在公式 3.24 中，$\hat{r}$ 是随机变量，总和 T_z 是常数。$\hat{t}_{rat}$的方差可以简写成 $V_{srs}(\hat{t}_{rat}) = T_z^2 \times V_{srs}(\hat{r})$。这里引入比率的估计值 $\hat{r}$ 的 SRSWOR 的设计方差(公式 2.9)，总和的比率估计值的近似方差为：

$$V_{srs}(\hat{t}_{rat}) \approx N^2\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right) \sum_{k=1}^{N} \frac{(Y_k - R \times Z_k)^2}{N - 1}, \tag{3.25}$$

其估计值为：

$$\hat{v}_{srs}(\hat{t}_{rat}) = N^2\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right) \sum_{k=1}^{n} \frac{(y_k - \hat{r}z_k)^2}{n - 1}。 \tag{3.26}$$

通过考察方差公式 3.25 中的平方和，能够找出比率估算提高总和估计效率的条件。总平方和可以分解成：

$$\begin{aligned}\sum_{k=1}^{N} (Y_k - R \times Z_k)^2/(N - 1) &= \sum_{k=1}^{N} [(Y_k - \bar{Y}) - R(Z_k - \bar{Z})]^2/(N - 1) \\ &= \sum_{k=1}^{N} [(Y_k - \bar{Y})^2 - R^2(Z_k - \bar{Z})^2 - \\ &\quad 2R(Y_k - \bar{Y})(Z_k - \bar{Z})]/(N - 1) \\ &= S_y^2 + R^2 S_z^2 - 2R\rho_{yz} S_y S_z,\end{aligned}$$

其中，ρ_{yz}是变量 y 与 z 的有限总体相关系数。考虑到下面的差，

$$V_{srs}(\hat{t}) - V_{srs}(\hat{t}_{rat}) = N^2\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)[S_y^2 - (S_y^2 + R^2 S_z^2 - 2R\rho_{yz} S_y S_z)]。$$

当 $V_{srs}(\hat{t}) > V_{srs}(\hat{t}_{rat})$时，比率估算提高了效率，即是，

$$R^2 S_z^2 < 2R\rho_{yz} S_z S_y$$

成立，或是

$$2\rho_{yz} > \frac{RS_z}{S_y}。$$

应当指出，$R=\overline{Y}/\overline{Z}$。前一个条件用变量 z 与 y 间的离差系数(C. V)来表示，则有：

$$\rho_{yz}>\left(\frac{1}{2}\right)\frac{\mathrm{C.V}_y}{\mathrm{C.V}_z},$$

其中，$\mathrm{C.V}_y=S_y/\overline{Y}$ 与 $\mathrm{C.V}_z=S_z/\overline{Z}$ 分别是 y 与 z 的离差系数。因此，效率上的提高取决于研究变量与辅助变量 y 与 z 以及它们各自的 C. V。

范例 3.11

1991 年省级人口比率估算总和的效率。变量 UE91 是研究变量 y，HOU85 则为辅助变量 z。UE91 与 HOU85 的相关系数为 $\rho_{yz}=0.996\ 7$。而相应的离差系数为，$\mathrm{C.V}_y=S_y/\overline{Y}=743/472=1.57$ 与 $\mathrm{C.V}_z=S_z/\overline{Z}=4\ 772/2\ 867=1.66$。上述给出的条件是有效的，因为：

$$\rho_{yz}=0.996\ 7>0.472\ 9=\frac{1}{2}\times\frac{1.57}{1.66}。$$

可以看出，比率估算提高了效率。这样的提高也可以被直接当成设计效应。除了给出的参数外，还需要比率 $R=\overline{Y}/\overline{Z}=472/2\ 867=0.164\ 6$。1991 年省级人口中总和的比率估计 $\hat{t}_{rat}$ 设计效应的数值为：

$$\begin{aligned}\mathrm{DEFF}_{srs}(\hat{t}_{rat})&=\frac{S_y^2+R^2S_z^2-2R\rho_{yz}S_yS_z}{S_y^2}\\&=\frac{743^2+0.164\ 6^2\times4\ 772^2-2\times0.164\ 6\times0.996\ 7\times743\times4\ 772}{743^2}\\&=0.010\ 2\end{aligned}$$

它接近于 0。效率的这一可观的提高，是由于 UE91 与 HOU85 间有利的关系，造成了比率 Y_k/Z_k 在总体中近乎于常数。

在实际中，比率估算的总和是通过使用已有的抽样设计的调查数据。如果使用了分层 SRS 的设计，则相应的参数将使用合适的层级权重来估算。目前的例子使用了无放回式简单随机抽样。下一个例子中也将使用这一设计。另外，其中也将演示 g 权重的使用。

范例 3.12

1991 年省级人口中用无放回式简单随机抽样情形下总和的比率估算。我们再次使用 UE91 为研究变量，HOU85 为辅助变量。从表 3.15 中的样本中，估算的比率 $\hat{r}=\bar{y}/\bar{z}=0.160\ 3$。样本标签是 STR = 1，ID 是整群标签，而权重为 WEIGHT = 4。

表 3.15　为比率估算准备的 1991 年省级人口中的一个 SRSWOR 样本

样本设计标识			元素标签	研究变量	辅助变量	g	Adj.
STR	CLU	WGHT		UE91	HOU85	WGHT	WGHT
1	1	4	Jyväskylä	4 123	26 881	0.556 2	2.224 8
1	4	4	Keuruu	760	4 896	0.556 2	2.224 8
1	5	4	Saarijärvi	721	3 730	0.556 2	2.224 8
1	15	4	Konginkangas	142	556	0.556 2	2.224 8
1	18	4	Kuhmoinen	187	1 463	0.556 2	2.224 8
1	26	4	Pihtipudas	331	1 946	0.556 2	2.224 8
1	30	4	Toivakka	127	834	0.556 2	2.224 8
1	31	4	Uurainen	219	932	0.556 2	2.224 8

抽样比例:8/32 =0.25。

为了估算总和,首先要计算标准权重 w_k^*。与前面一样,抽样权重 w_k 是一常数,$w_k = N/n = 32/8 = 4$。g 权重的数值为 $g_k = T_z/\hat{t}_z$。辅助变量的总体总和为 $T_z = 91\ 753$,而其从样本中计算的估计值为 $\hat{t}_z = 164\ 952$。所以,常数 g 权重为,$g_k = 91\ 753/164\ 952 = 0.556\ 2$。用 g 权重乘以 w_k 得到标准权重$w_k^* = 4 \times 0.556\ 2 = 2.224\ 8$。

总和的变量估计值为:

$$\hat{t}_{rat} = \sum_{k=1}^{n} w_k^* y_k = \hat{r} \times T_z = 0.160\ 3 \times 91\ 753 = 14\ 707\ ,$$

与 SRSWOR 下的失业总人数估计值 $\hat{t} = 26\ 440$ 相比,它非常接近于总体总和 $T = 15\ 098$。总和估计值的方差估计值为:

$$\hat{v}_{srs}(\hat{t}_{rat}) = 32^2 \times \frac{1 - 0.25}{8} \times 91^2 = 892^2。$$

相应的 deff 估计值为:

$$\text{deff}_{srs}(\hat{t}_{rat}) = \frac{\hat{v}_{srs}(\hat{t}_{rat})}{\hat{v}_{srs}(\hat{t})} = 892^2/13\ 282^2 = 0.004\ 5\ ,$$

这也显示了,比率估算提高了效率。总体总和 T_z 的最少信息以及 z 的样本取值产出了很好的结果。

也可以使用比率估算的总和来计算 DEFF。这是因为,方差 $V_{srs}(\hat{t}_{rat})$ 为,

$$V_{srs}(\hat{t}_{rat}) \approx N^2\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)\sum_{k=1}^{N}\frac{(Y_k - R \times Z_k)^2}{N - 1}$$

$$= 32^2 \times \frac{(1 - 0.25)}{8} \times 75^2 = 736^2。$$

除以相应的 $\hat{t}$ 的 SRSWOR 方差,有,

$$\mathrm{DEFF}_{srs}(\hat{t}_{rat})=\frac{V_{srs}(\hat{t}_{rat})}{V_{srs}(N\bar{y})}=736^2/7\ 283^2=0.010\ 2,$$

这与范例3.11中的数值相同。

从这些数据可以看出,变量估算极大地提高了效率,并使得总和的点估计更接近于总体真实值。比率估计值的取值取决于比率 Y_k/Z_k 在总体是常数。应当注意到,即使是两变量间的高相关也并不能保证这一点。这是因为,比率估算假定了 y 与 z 之间的回归直线接近原点。因此,相应的回归方程中没有截距项。即使是相关系数并不接近于0,当总体回归直线与 y 轴的截距远离原点时,比率估算也许就并不合适。这种情形下,下面的方法更恰当些。

总和的回归估算

研究变量 y 的总体总和 T 的回归估算是基于 y 与连续变量 z 之间的线性回归。这一线性回归可以是,方差为 $V_M(y_k)=\sigma^2$ 的 $E_M(y_k)=\alpha+\beta\times z_k$。其中,$y_k$ 是实现值为总体取值 Y_k 的随机自变量,α,β,σ^2 为未知参数,Z_k 是已知总体中 z 的取值,E_M 与 V_M 分别是模型的期望值与方差。有限总体中对应 α,β 的 A 与 B 是从样本中通过加权最小二乘法来估算的。这样的做法考虑到了抽样设计。很快注意到的是,可以纳入多个辅助变量到模型之中。注意,模型的假设中引入了一种新的随机性,在前面的估算方法中,样本选取是随机变化的唯一来源。

我们将使用上述只有一个辅助变量的回归模型,来讨论无放回式SRS情形下回归估算的基本原则。使用最小二乘法来估算有限总体的数值 A 与 B,得出斜率 B 的估计值 $\hat{b}=\hat{s}_{yz}/\hat{s}_z^2$,截距 A 的估计值 $\hat{a}=\bar{y}-\hat{b}\,\bar{z}$。使用估计值 $\hat{b}$,y 的总和 T 的回归估计值为,

$$\hat{t}_{reg}=N[\bar{y}+\hat{b}(\bar{Z}-\bar{z})]=\hat{t}+\hat{b}(T_z-\hat{t}_z) \tag{3.27}$$

其中,$\hat{t}=N\bar{y}$ 是 T 的SRSWOR估计值,$\hat{t}_z=N\bar{z}$ 是 T_z 的SRSWOR估计值,$\bar{Z}=T_z/N$。如果使用转换值 $z_k^*=\bar{Z}-z_k$ 而非 z_k,模型中截距的估计值为 $\hat{a}^*=\hat{a}+\hat{b}\bar{Z}$,得出 $\hat{t}_{reg}=N\hat{a}^*$。这是因为,式3.27也可以写成 $\hat{t}_{reg}=N\hat{a}+\hat{b}T_z$。注意,总和 T 的估算只假定已知总体总和 T_z 及辅助变量 z 的样本取值 z_k。

回归估算包括一系列估算公式。比如,前面的比率估算 $\hat{t}_{rat}=\hat{r}T_z$ 可以看成是式3.27的特例,其中的截距 A 被假定为0,斜率 B 为 $\hat{b}=\hat{r}=\hat{t}/\hat{t}_z$。

另外,我们可以计算标准权重 $w_k^*=w_k\times g_k$。其中,w_k 是抽样权重,g 权重的计算是,

$$g_k=\frac{N}{\hat{N}}\left[1+\frac{\bar{Z}-\bar{z}}{\frac{n-1}{n}\hat{s}_z^2}\times(z_k-\bar{z})\right],$$

其中,$\bar{Z}$ 是辅助变量 z 的总体均值,$\bar{z}$ 是其样本均值,抽样权重的总和为 $\sum_{k=1}^{n}w_k=$

$\hat{N}$,而

$$\hat{s}_z^2 = \frac{\sum_{k=1}^{n}(z_k - \bar{z})^2}{n-1}。$$

表3.16 给出了在SRSWOR 情形下,1991 年省级人口的模型 $E_M(y_k) = \alpha + \beta \times z_k$ 的权重 g_k 及标准化权重 w_k^*。总体总和的回归估计即是,标准化权重 w_k^* 乘以观测值 y_k,并对于所有样本元素加总。公式3.27 中的回归估算也可以表示为 $\hat{t}_{reg} = \sum_{k=1}^{n} w_k^* y_k$。

在 SRSWOR 情形下,$\hat{t}_{reg}$的近似设计方差为,

$$V_{srs}(\hat{t}_{reg}) \approx N^2\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)S_E^2, \tag{3.28}$$

其中,$S_E^2 = \sum_{k=1}^{N}(E_k - \bar{E})^2/(N-1)$,$E_k = Y_k - \hat{Y}_k$,而 $\bar{E} = \sum_{k=1}^{N} E_k/N$ 是总体残差的均值。拟合值 $\hat{Y}_k = A + B \times Z_k$ 从总体取值中计算得出。在 SRSWOR 情形下,用 $\hat{s}_e^2 = \sum_{k=1}^{n}(\hat{e}_k - \bar{\hat{e}})^2/(n-1)$ 替换 S_E^2,可以计算出 $\hat{t}_{reg}$的近似设计方差估计值,其中,$\hat{e}_k = y_k - \hat{y}_k$,$\bar{\hat{e}} = \sum_{k=1}^{n}\hat{e}_k/n$。拟合值 $\hat{y}_k = \hat{a} + \hat{b} \times z_k$ 从样本取值中计算得出。另一个使用 g 权重的保守的估计值为,

$$\hat{v}_{srs}(\hat{t}_{reg}) = N^2\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)\left(\frac{n-1}{n-p}\right) \times \hat{s}_{\hat{e}^*}^2, \tag{3.29}$$

其中,$\hat{s}_{\hat{e}}^2 = \sum_{k=1}^{n}(\hat{e}_k^* - \bar{\hat{e}}^*)^2/(n-1)$,$\hat{e}_k^* = g_k \times \hat{e}_k$,$\bar{\hat{e}}^* = \sum_{k=1}^{n}\hat{e}_k^*/n$,而 p 是模型中估算的参数数目。

与相应的简单随机抽样的估计值相比,回归估算所获得的效率取决于有限总体中变量 y 与 z 间的相关系数$\rho_{yz} = S_{yz}/(S_yS_z)$的大小。改写式3.28 中的近似方差的形式,可以看得更清楚,

$$V_{srs}(\hat{t}_{reg}) \approx N^2\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)S_y^2(1 - \rho_{yz}^2)。 \tag{3.30}$$

下面要指出,相关系数的大小对于回归估算提高的效率有着决定性的作用。当ρ_{yz}为0 时,回归估算 $\hat{t}_{reg}$的方差与 SRSWOR 估算 $\hat{t}$ 的方差相等。但相关系数不为0 时,方差明显降低。

在一定条件下,总和的回归估算与比率估算相比更为有效。下面将演示SRSWOR、比率估算及回归估算的方差。使用了无放回式简单随机抽样,公式中的常数 c 表示 $c = N^2 \times [1 - (n/N)] \times (1/n)$。各个方差为,

基于设计的估算公式	$V_{srs}(\hat{t}) = cS_y^2$
比率估算公式	$V_{srs}(\hat{t}_{rat}) = c(S_y^2 + R^2S_z^2 - 2R\rho_{yz}S_yS_z)$
回归估算公式	$V_{srs}(\hat{t}_{reg}) = cS_y^2(1 - \rho_{yz}^2)$

考察回归系数 B 与比率 $R=T/T_z$ 之间的关系,将揭示在什么条件下回归估算的总和比比率估算的总和更为有效。为了找出这一条件,两个方差间的差为,

$$V_{srs}(\hat{t}_{rat})-V_{srs}(\hat{t}_{reg})=c[(S_y^2+R^2S_z^2-2R\rho_{yz}S_yS_z)-S_y^2+\rho_{yz}^2S_y^2]$$
$$=c[(R^2S_z^2-2R\rho_{yz}S_yS_z)+\rho_{yz}^2S_y^2]。$$

当差为正时,回归估算更为有效:

$$R^2S_z^2-2R\rho_{yz}S_yS_z+\rho_{yz}^2S_y^2>0$$

这一条件可以写成,

$$-\rho_{yz}^2S_{y-}^2<R^2S_z^2-2R\rho_{yz}S_yS_z。$$

上述不等式除以 $S_z^{\,2}$,并代入 $\rho_{yz}=S_{yz}/(S_yS_z)$ 与 $B=S_{yz}/S_z^{\,2}$,有,

$$-B^2<R^2-2RB。$$

因而,回归估算比比率估算更为有效,当

$$(B-R)^2>0。$$

因而,有限总体的回归系数与比率之间的差的平方决定回归估算是否更为有效。

回归估算也可以使用多元回归模型作为辅助模型。我们假定研究变量 y 与 p 个连续辅助变量 $z_1,z_2,\cdots,z_p$ 的线性回归模型为,$y_k=\alpha+\beta_1z_{1k}+\beta_2z_{2k}+\cdots+\beta_pz_{pk}+\varepsilon_k$。其中,$\alpha$ 表示截距,$\beta_j(j=1,\cdots,p)$ 是斜率参数,而 ε_k 是残差。对于多元回归估算,我们假定已知各个辅助变量的总体总和 $T_{z_1},T_{z_2},\cdots,T_{z_p}$。可以由调查以外的来源获得这些数据,比如出版的官方统计。y 的总体总和 T 的回归估计值为,

$$\hat{t}_{reg}=\hat{t}+\hat{b}_1(T_{z1}-\hat{t}_{z1})+\hat{b}_2(T_{z_2}-\hat{t}_{z_2})+\cdots+\hat{b}_p(T_{z_p}-\hat{t}_{z_p}),\qquad(3.31)$$

其中,估算的回归系数 $\hat{b}_1,\hat{b}_2,\cdots,\hat{b}_p$ 是从样本数据中使用加权最小二乘法获得,权重 $w_k=1/\pi_k$。估计值 $\hat{t}$ 与 $\hat{t}_{z_j}(j=1,\cdots,p)$ 是霍维茨-汤普森估计值。

通常被称为通用回归(GREG)估算值的另外一个形式(Sarndal, et al., 1992)为,

$$\hat{t}_{reg}=\sum_{k=1}^{N}\hat{y}_k+\sum_{k=1}^{n}w_k(y_k-\hat{y}_k),\qquad(3.32)$$

其中,$\hat{y}_k=\hat{a}+\hat{b}_1z_{1k}+\hat{b}_2z_{2k}+\cdots+\hat{b}_pz_{pk}$是使用估算的回归系数及已知的 z 值得出的拟合值。注意式 3.31 与 3.32 间的差别。在前者中,我们只需要知道辅助变量 z 的总体总和,而在后者中,假定已知变量 z 在每一个总体元素的取值(因为第一个加总是对于所有总体元素的)。因此,式 3.32 比式 3.31 需要总体更详细的信息。在统计基础较好的地方,从行政登记中得来的人口普查或是相似的统计登记被用作抽样框架。这时候,辅助变量 z 的微观层次的数据可能的确存在。在这样的例子中,总体框架通常包含必需的辅助变量 z 的微观层次的数据(参见第 6 章)。

让我们进一步考察多元回归估计值在式 3.32 中的表示法。显然,当所有样本元素的权重相等时,含有截距的模型使用最小二乘法,式 3.32 中的后半

部分就没有了，而回归估计值减少为总体中拟合值的加总。诸如简单随机抽样这样的自我加权的设计就是如此。但是，当各元素的权重各不相同时，加权残差总和可能不等于 0。非比例配额的分层 SRS 的例子就是如此。在这样的例子中，式 3.32 中的后半部分是避免假定错误的偏差校正因子。

在 SRSWOR 情形下，式 3.28 中给出的设计方差的近似值可以使用拟合值 $\hat{Y}_k = A + B_1 Z_{1k} + \cdots + B_p Z_{pk}$ 来得出。用 $\hat{y}_k = \hat{a} + \hat{b}_1 z_{1k} + \hat{b}_2 z_{2k} + \cdots + \hat{b}_p z_{pk}$ 来代替 $\hat{Y}_k$，可以求出方差估计值。另一个方差估计值的计算为，

$$\hat{v}_{srs}(\hat{t}_{reg}) = \hat{v}_{srs}(\hat{t})(1 - \hat{R}^2), \tag{3.33}$$

其中，多元相关系数的平方 $\hat{R}^2$ 从样本数据中计算得出。由于这一项总是非负数，多元回归估计值总是至少与无放回式简单随机抽样同等有效。当与研究变量 y 相关的辅助变量 z 的数据被纳入估算过程时，效率将得到提高。

在下一个例子中，我们从样本数据中，首先计算单个辅助变量的回归估算的总和，然后将这一过程应用于多元回归估算中。

范例 3.13

单一辅助变量

1991 年省级人口数据中总和的回归估算。我们使用前面抽取的简单随机样本。用辅助变量 HOU85 来回归研究变量 UE91。我们使用两种方法进行回归估算，估计值相同。首先将 HOU85 用作预测变量，使用估算的截距 $\hat{b}$ 计算估计值 $\hat{t}_{reg}$。表 3.16 中，样本标签与 SRSWOR 例子相同。抽样比例与前面的相同，也是 0.25。

表 3.16 为回归估算准备的 1991 年省级人口中的一个 SRSWOR 样本

样本设计标识			元素标签	研究变量	辅助信息			
					变 量	模型组	权 重	
STR	CLU	WGHT		UE91	HOU85		g-权重	w^*-权重
1	1	4	Jyväskylä	4 123	26 881	1	0.284 4	1.137 8
1	4	4	Keuruu	760	4 896	1	1.008 5	4.034 1
1	5	4	Saarijärvi	721	3 730	1	1.046 9	4.187 7
1	15	4	Konginkangas	142	556	1	1.105 7	4.605 8
1	18	4	Kuhmoinen	187	1 463	1	1.121 6	4.486 3
1	26	4	Pihtipudas	331	1 946	1	1.139 1	4.422 7
1	30	4	Toivakka	127	834	1	1.142 3	4.569 1
1	31	4	Uurainen	219	932	1	1.151 5	4.556 2

抽样比例 = 8/32 = 0.25。

UE91 为因变量，HOU85 为预测变量，估算的斜率为 $\hat{b} = 0.152$，得出，

$\hat{t}_{reg} = \hat{t} + \hat{b}(T_z - \hat{t}_z) = 26\ 440 + 0.152 \times (91\ 753 - 164\ 952) = 15\ 312$。

使用标准权重，可以获得相同的点估计，$\hat{t}_{reg} = \sum_{k=1}^{8} w_k^* y_k = 15\ 312$（见表3.16）。式3.29或式3.33可以用来估算方差。特别是在本例子这样的样本规模较小时，前者给出保守的估计。因而，使用式3.29，我们有，

$$\hat{v}_{srs}(\hat{t}_{reg}) = N^2\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)\left(\frac{n-1}{n-p}\right) \times \hat{s}^2_{\hat{e}}。$$

$$= 32^2 \times \left(1 - \frac{8}{32}\right) \times \left(\frac{8-1}{8-2}\right) \times \left(\frac{1}{8}\right) \times 61.24^2 = 648^2。$$

相应的SRSWOR情形下，基于设计的总和估计值为 $\hat{t} = 26\ 440$，其标准误为13 282。deff的估计值为 deff $= 648^2/13\ 282^2 = 0.002$，接近于0。这是回归估算在当前估算问题中，优于基于设计的估算的令人信服的证据。效率提高的原因是UE91与HOU85之间的高度线性关系。

多元回归模型

1991年省级人口数据中总和的多元回归估算。用HOU85与URB85（城镇取值为1，其余为0；见表2.1）来回归研究变量UE91。我们使用公式3.31及公式3.32中的GREG方法。首先，通过将表3.16中含有 $n=8$ 个自治市样本数据拟合两个预测变量的回归模型，估算出回归系数 $\hat{b}_1$ 与 $\hat{b}_2$。估计值为，$\hat{b}_1 = 0.149\ 56$，$\hat{b}_2 = 68.107$。与前面相同，辅助变量的估算总和 $\hat{t}_{z_1} = 164\ 952$，$\hat{t}_{z_2} = 12$。同时，我们知道总体总和 $T_{z_1} = 91\ 753$，$T_{z_2} = 7$。使用式3.31我们有，

$\hat{t}_{reg} = \hat{t} + \hat{b}_1(T_{z_1} - \hat{t}_{z_1}) + \hat{b}_2(T_{z_2} - \hat{t}_{z_2}) = 26\ 440 + 0.149\ 56 \times (91\ 753 - 164\ 952) + 68.107 \times (7 - 12) = 15\ 152$。

使用式3.32，我们首先计算所有总体元素的拟合值。总体中拟合值的总和即是想要的回归估计值。表3.17简单列出了GREG估算过程。这里，也得出了估计值15 152。注意，在SRSWOR情形下，抽样权重相同，设定样本数据中的残差和为0。

在计算了样本数据的多元相关系数平方 $\hat{R}^2 = 0.998$ 之后，由式3.33得出方差估计值，$\hat{v}(\hat{t}_{reg}) = 569^2$。它小于前面只有HOU85作为辅助变量的例子，其中的估计值为 $\hat{v}(\hat{t}_{reg}) = 648^2$。因此，在这里，多元回归估算看起来略微高效。这里的设计效应估计值为 deff $= 569^2/13\ 282^2 = 0.001\ 8$。

表 3.17 为多元回归估算加入样本数据的总体框

		总体框			样本		模型拟合	
序号 k	标签	URB85 z_{1k}	HOU85 z_{2k}	样本标签	权重 w_k	UE91 y_k	拟合值 $\hat{y}_k$	残差 $\hat{e}_k$
1	**Jyväskylä**	**1**	**26 881**	**1**	**4**	**4 123**	**4 118.15**	**4.85**
2	Jämsä	1	4 663	0	…	…	795.27	…
3	Jämsänkoski	1	3 019	0	…	…	549.40	…
4	**Keuruu**	**1**	**4 896**	**1**	**4**	**760**	**830.12**	**−70.12**
5	**Saarijärvi**	**1**	**3 730**	**1**	**4**	**721**	**655.73**	**65.27**
6	Suolahti	1	2 389	0	…	…	455.18	…
7	Äänekoski	1	4 264	0	…	…	735.60	…
8	Hankasalmi	0	2 179	0	…	…	355.66	…
9	Joutsa	0	1 823	0	…	…	302.42	…
10	J:kylä mlk	0	9 230	0	…	…	1 410.20	…
11	Kannonkoski	0	726	0	…	…	138.36	…
12	Karstula	0	1 868	0	…	…	309.15	…
13	Kinnula	0	675	0	…	…	130.73	…
14	Kivijärvi	0	634	0	…	…	124.60	…
15	**Konginkangas**	**0**	**556**	**1**	**4**	**142**	**112.93**	**29.07**
16	Konnevesi	0	1 215	0	…	…	211.49	…
17	Korpilahti	0	1 793	0	…	…	297.93	…
18	**Kuhmoinen**	**0**	**1 463**	**1**	**4**	**187**	**248.58**	**−61.58**
19	Kyyjärvi	0	672	0	…	…	130.28	…
20	Laukaa	0	4 952	0	…	…	770.39	…
21	Leivonmäki	0	545	0	…	…	111.29	…
22	Luhanka	0	435	0	…	…	94.83	…
23	Multia	0	925	0	…	…	168.12	…
24	Muurame	0	1 853	0	…	…	306.91	…
25	Petäjävesi	0	1 352	0	…	…	231.98	…
26	**Pihtipudas**	**0**	**1 946**	**1**	**4**	**331**	**320.82**	**10.18**
27	Pylkönmäki	0	473	0	…	…	100.52	…
28	Sumiainen	0	485	0	…	…	102.31	…
29	Säynätsalo	0	1 226	0	…	…	213.13	…
30	**Toivakka**	**0**	**834**	**1**	**4**	**127**	**154.51**	**−27.51**
31	**Uurainen**	**0**	**932**	**1**	**4**	**219**	**169.16**	**49.84**
32	Viitasaari	0	3 119	0	…	…	496.25	…
	总和	**7**	**91 753**	**8**	**32**	**6 610**	**15 151.98**	**0.00**

数据采自 1991 年省级人口中的一个 SRSWOR 样本；

“…”未抽中元素。

我们用设定 SRSWOR、含有一个或是两个辅助变量的简单例子来演示了回归估算。这一方法可以用于更为复杂的设计中。同时,多个辅助变量也可以纳入估算过程中。而加权最小二乘法也可用于此。虽然,回归估算中使用多变量回归模型在技术上直截了当,但与简单随机抽样的回归估算相比,它含有一定的复杂性,比如预测变量之间的多元共线性。由于非连续性变量可以纳入到回归模型中,另一个扩展显而易见。在回归估算中使用这类辅助变量,将得出类似于方差分析的模型。第 6 章将讨论与次级总体相连接的进一步的扩展。

估算策略的比较

对于模型辅助估算,我们设立了三组新的权重,用 w^* 表示。首先,我们考察了这些权重的标准化特征。在比率估算中,辅助变量 z 的标准化方程为,

$$\sum_{k=1}^{n} w_k^* \times z_k = T_z$$

其中,$T_z = \sum_{k=1}^{N} Z_k = 91\ 753$。在回归估算中也是如此。

前面从 1991 年省级人口数据中,以 SRSWOR 抽取了样本。接下来,我们比较由这一样本得来的模型辅助估算结果。更具体的,比较了 UE91 总体总和 T 的后续分层、比率估算以及回归估算的结果。以标准 SRS 公式得来的基于设计的估计值也包含在其中(见表 3.18)。已知的 UE91 总体总和 $T=15\ 098$作为参考数字。

表 3.18 各种估算策略下 UE91 总体总和的估计值

估算策略		估算量	估计值	s. e	deff
		基于设计的			
SRSWOR		$\hat{t}_{srswor}$	26 440	13 282	1.000 0
SRSWR		$\hat{t}_{srswr}$	26 440	15 095	1.291 7
		基于设计模型辅助的			
后续分层估计值		$\hat{t}_{pos}$	18 106	6 021	0.332 3
比率估计值		$\hat{t}_{rat}$	14 707	892	0.004 5
回归估计值	一个 z-变量	$\hat{t}_{reg,1}$	15 312	648	0.002 0
	两个 z-变量	$\hat{t}_{reg,2}$	15 152	569	0.001 8

数据采自从 1991 年省级人口抽取的 8 个元素的 SRSWOR 样本。

可以很容易地得出两个结论。第一,使用辅助信息计算的点估计比基于

设计的估计值更接近于总体总和。第二,模型辅助的估计值比 SRSWOR 高效得多。

后续分层估计值使用了非连续的辅助信息,自治市的城镇与农村行政划分。因为失业数字的离异在后续层级中小于总体中的离异,与研究变量的这层关系导致了估计值的高效。但是,这样的关系并没有 UE91 与连续的家庭数目 HOU85 间的关系强烈。可以从比率与回归估算结果中看出这一特点。由于比率估算假定 UE91 与 HOU85 的回归直线经过原点,但这又并非事实,所以回归估算比比率估算略微高效。

小 结

在研究变量与辅助变量有着较强关系的情形下,当估算有限总体的参数时,使用总体中的辅助信息是获取更为准确估计值的有力工具。真是如此的话,有效估计值可以产生接近于真实总体数值以及较小的标准误。辅助变量可以是非连续性的,它使用于后续分层。如果辅助变量是连续性的,则适用于比率与回归估算。

模型辅助估算通常用于描述性调查中,以提高对感兴趣的研究变量的总体总和的估算。但在研究变量数目更多的多重目的研究中,可能很难找到满足这一目的的辅助变量。在这样的调查中,后续分层经常被用来校正无应答的情况。

借助计算演示,我们在这里考察了模型辅助估算的基础原则。若需要更多的详情,读者可以参阅桑德尔等(Särndal et al.,1992),那里广泛地讨论了涵盖后续分层、比率估算以及回归估算的模型辅助调查抽样。这些方法被看成是扩展回归估算的特例。许多统计机构使用扩展回归估算来生成官方统计,例如埃斯特佛奥等(Estevao et al. 1995)。霍尔特与史密斯(Holt and Smith,1979)给出了一个清楚的后续分层的总览。作为后续分层的扩展,德维尔与桑德尔(Deville and Särndal,1992)以及德维尔等(Deville et al.,1993)讨论了一系列对于已知边缘总和标准化的权重。席尔瓦与斯金纳(Silva and Skinner,1997)讨论了回归估算中的变量选择问题。

3.4 使用设计效应比较效率

在不同的抽样设计中,设计效应是一个比较估算总体参数的便利工具。在这一小节中,我们将总结前面小节中效率评估的结论。

效率通过比较估计值与 SRSWOR 得到的估计值的方差而获得。它是由总体的设计效应(DEFF)或是从样本数据中计算出来的设计效应来表示的。

前面,我们通过3种方法来评估效率:(1)分析性的,推导相应的设计方差公式;(2)基于总体的,计算小规模固定总体——1991年省级人口数据——设计方差的真实值;(3)基于样本的,从应用于1991年省级人口数据的抽样设计中,估算出设计方差。对这些方法的评估涵盖了所有讨论过的基本抽样技术。在使用估算的deff、基于样本的设计效应的评估中,我们讨论了总和、比率以及中位值估计值。

让我们首先考虑对研究变量y的总和T的估算的效率评估。设计效应的定义是两个设计方差的比率:恰当反映抽样设计的、总和估计值$\hat{t}^*$的真实方差$V_{p(s)}(\hat{t}^*)$,以及从SRSWOR设定中导出的方差$V_{srs}(N\bar{y})$。其中,$\hat{t}^*$是在设计$p(s)$情形下,基于设计的总和的估计值;$N\bar{y}=\hat{t}$是相应的SRSWOR估计值。注意,总和的这两个估计值可以不同,而对于实际的抽样设计,设定相同的样本规模。所以,如小节2.1中定义的,DEFF为,

$$\text{DEFF}_{p(s)}(\hat{t}^*)=\frac{V_{p(s)}(\hat{t}^*)}{V_{srs}(N\bar{y})} \tag{3.34}$$

这一等式表示,当DEFF>1,实际设计比SRSWOR低效;当DEFF接近于1,这两个设计效率相等;当DEFF<1,则实际设计比SRSWOR更为有效。

设计效应的分析性评估

当方差等式中的总体参数——如总体方差S^2——在设计效应的公式中抵消掉,则DEFF的分析性评估成为可能。比如,对于给定的样本规模n和总体N,可以计算放回式简单随机抽样(SRSWR)的设计效应。因此,我们有DEFF$=(N-1)/(N-n)$,SRSWR的设计效应大于或是等于1。有时,也可以找出DEFF小于1,实际设计比SRSWOR更为有效的条件。

在简单随机分层抽样、概率对应某一规模变量的比例抽样以及整群抽样中,总和估计值的设计效应的分析性评估将得以演示。由于系统抽样可以被看成是整群抽样的特例,所以没有包括它。

1. 按比例配额的分层抽样(STR) 影响STR的效率的因素是不同层级的异质性以及层级内的同质性。研究变量y的总和T的估计值$\hat{t}^*=\hat{t}$的设计效应为,

$$\text{DEFF}_{str}(\hat{t})\approx\frac{\sum_{h=1}^{H}W_hS_h^2}{S^2}, \tag{3.35}$$

   其中,S_h^2是层级方差,S^2是y的总体方差(见章节3.1)。在分层抽样中,DEFF通常小于1。这种情况发生的情形是,相对于研究变量的离异,层级内部同质性高,亦即层内方差较小。

2. 概率对应规模变量的抽样(PPS) 这里需要度量总体元素规模的辅助变量z对于总体中所有单位的取值。假定y与z的总体回归直线与y

轴的交点接近于原点,估计值 $\hat{t}=\hat{t}_{HT}$的设计效应的近似等式为,

$$\mathrm{DEFF}_{pps}(\hat{t}_{HT})\approx(1-\rho_{yz}^{2}), \tag{3.36}$$

其中,ρ_{yz}是有限总体中研究变量 y 与规模变量 z 之间的相关系数(见章节2.5)。给定上面的条件,如果 z 是一个合适的规模变量,与 y 强相关,则得到小于 1 的 DEFF。

3. 整群抽样(CLU) 在 CLU 情形下的设计效应取决于衡量总体整群同质性的研究变量 y 的群内系数 ρ_{int}。设定等规模整群,估计值 $\hat{t}$ 的设计效应的近似等式为,

$$\mathrm{DEFF}_{clu}(\hat{t})\approx 1+(B-1)\rho_{int} \tag{3.37}$$

其中,B 是整群规模(见章节 3.2)。由于在整群抽样中,整群通常内部同质性高,导致正的 ρ_{int},因而设计效应倾向于大于 1。

为了在抽样设计中充分利用上述公式,有必要了解总体中研究变量的离异。在选择一个抽样设计时,计划者还需要知道在层级和整群层次上的离异情况,以及研究变量与规模变量间的相关信息。但在实际中,很少现成的这样的信息。但是,在某些情形下,从其他辅助来源或是小型的前期试探性调查中,可以获得近似的信息。

总体设计效应

我们接下来,使用从 1991 年省级人口数据中 6 个样本设计相应的公式来计算设计方差,并以此来评估总和的总体设计效应。从总体 32 个自治市中($N=32$)抽取的固定的样本规模为 8($n=8$)。表 3. 19 给出了总体设计效应的数值。

表 3. 19 1991 年省级人口中各种抽样设计(固定样本规模 $n=8$)下总和估计量的总体 DEFF

抽样设计		S. E	DEFF
对应规模抽样(无放回)	PPS	720	0. 01
分层抽样(指数配额)	STR	4 852	0. 44
系统抽样(随机起点)	SYS	5 420	0. 55
整群抽样(两级)	CLU2	6 532	0. 80
整群抽样(一级)	CLU1	6 663	0. 84
简单随机抽样	SRSWOR	7 283	1. 00

看起来,在估算总和时,概率对应规模的 PPS 抽样是最为有效的抽样设计。其总体 DEFF 极小,为 0. 01。效率提高的原因是,UE91 与 HOU85(作为规模变量)间其比率在总体中几乎为常数的关系。应该注意到,研究变量 UE91 在总体中的分布也影响效率。1991 年省级人口数据中的 UE91 的分布

极为偏斜。但是,在 PPS 中,较大的选中概率给了较大的整群,因而使得从总体中抽取的样本的构成变动较小。所以,样本总和不至于在各个样本间变动较大,这就导致了有效的估算。研究变量与规模变量间的强相关关系也有利于提高效率。在这个例子中,其相关系数接近于 1。

分层抽样的 DEFF 为 0.44,看起来在估算总和时也相当有效。但是 PPS 的优势还是相当明显。分层将自治市分为城镇和农村,城镇中的平均失业人数看起来要多于农村。这些层级看起来内部同质,这是一个提高效率的特征。由于抽样框架中有一个单调的趋势,组内相关系数接近于 0,导致较高的效率。两级整群抽样较 SYS 稍微低效,而一级整群抽样较两级整群抽样略微低效。

样本设计效应

前面的效率比较是理论上的,我们在总体层次上讨论了设计方差。接下来,我们根据从 1991 年省级人口数据中抽取规模 $n=8$ 的样本来评估效率。通过计算总体参数 θ 估计值相应的方差估计值 $\hat{v}_{p(s)}(\hat{\theta}^*)$ 与 $\hat{v}_{srs}(\hat{\theta})$,我们得到设计效应,

$$\text{deff}_{p(s)}(\hat{\theta}^*)=\frac{\hat{v}_{p(s)}(\hat{\theta}^*)}{\hat{v}_{srs}(\hat{\theta})}, \tag{3.38}$$

其中,$\hat{\theta}^*$ 是 θ 基于设计的估计值,$\hat{\theta}$ 是相应的 SRSWOR 中的估计值。

使用样本 deff(见表 3.20)比较了估计值 $\hat{t}^*$(总和)、$\hat{r}^*$(比率)与 $\hat{m}^*$(中位值)在各个抽样设计 $p(s)$ 下的样本中的估算效率。这些估计值在 1991 年省级人口数据中有着自然的意义。总和度量全省的失业人数(UE91),比率度量失业比例,中位值度量每自治市的平均失业人数。

表 3.20 1991 年省级人口中六种抽样设计下总和、比率与中位值估计值的样本设计效应

抽样设计		$\text{deff}(\hat{t}^*)$	$\text{deff}(\hat{r}^*)$	$\text{deff}(\hat{m}^*)$
对应规模抽样	PPS	0.003 5	0.19	0.92
分层抽样(指数配额)	STR	0.21	0.38	0.19
系统抽样(隐性分层)	SYS	0.76	1.29	0.21
整群抽样(两级)	CLU2	0.93	0.99	0.84
整群抽样(一级)	CLU1	1.92	1.44	1.29
简单随机抽样	SRSWOR	1.00	1.00	1.00

估计值 deff 不仅仅在不同的抽样设计中各不相同,同一设计中不同估计值间也不相同。PPS 与 STR 在估算总和时效率最高,因为其 deff 估计值接近于 0。对于比率而言,PPS 与 STR 优于其他设计,但其设计效应大于对于总和的估算。对于中位值,deff 估计值在含有隐性分层的 SYS 以及 STR 情形下接

近于0。

小 结

给定抽样设计，设计效应是评估估计值效率的实用工具。设计效应也可用来比较不同抽样设计间的效率。设计效应清楚地给出相当于简单随机抽样的复杂抽样的效应。取决于所讨论的估计值类型，抽样设计可以从几个方面影响到标量类型估计值的设计效应。总和估计值是线性类型，比率估计值是非线性类型，而中位值是均值的抗扰估计值。它们代表了统计分析中常用的估计值类型。重要的是，对于给定的估计值——比如总和——的最佳设计，将其标准误降至最低，亦即产生接近于0 的 deff 估计值，这样的最佳化标准并非一定要满足另一估计值。在我们的例子中，设计效应看起来并没有影响到中位值的估计值 $\hat{m}$。

在分析复杂调查数据时，可以成功地运用设计效应。在前面的几个小节中，我们使用设计效应的主要目的是描述性的，用它来解决小规模固定样本中的估算问题。在下面的章节中，我们将讨论分析性的情形，并给出更多的使用设计效应的实用例子。我们将讨论从大规模总体中抽取的复杂调查数据的估算和检验的问题。比如，我们将看到，使用设计效应（或是它的扩展）可以在适当地考虑到抽样设计的复杂性的情形下，估计标准误以及计算检验统计量的观测值。为了描述性和分析性的目的，商用软件可以计算出设计效应。同时，设计效应是度量计算中内在的复杂抽样效应的指标。这一问题更多的文献，参阅基什与弗兰克尔（Kish and Frankel，1974）以及基什（Kish，1995）的文章。

处理非抽样误差 第4章

Handling Nonsampling Errors

到现在为止讨论的调查估算方法中，离异的唯一来源是由估计值标准误衡量的抽样误差。除了抽样误差以外，调查中的离异也有其他来源造成的非抽样误差。这样的误差尤其存在于大规模调查中。调查机构努力在数据收集与数据处理阶段降低非抽样误差。完备的总体框架、周密计划和严格检验的测量工具、训练有素和态度端正的调查员及严格执行的实地工作和数据处理可以保证高应答率及测量与处理误差，以及一个总体上较好的调查质量。

非抽样误差的重要类型包括无应答、框架涵盖误差、测量误差以及数据处理误差。无应答表示并非所有的样本对于某一题器给出了回答。框架涵盖误差包括总体框架的缺陷。测量误差是指研究变量的观测值与真实值之间的差异。数据处理误差包括各种发生在将收集上来的数据转换成机器阅读形式过程中的误差，诸如数据录入、编码以及校正误差。

非抽样误差可以导致估算偏差。有多种方法来修正非抽样误差的恶果。在以下两个小节中，我们将详细讨论修正无应答造成的非抽样误差的方法。我们将通过使用前面描述过的方法来演示修正无应答。本章最后给出讨论总体调查质量的小结，也给出了更多的参考文献。

无应答

对于样本受访人，无法完全获得计划中的测量值或是回答被称作无应答。无应答造成缺损数据，亦即获得的研究变量 y 的数据小于计划中的规模。某一样本元素，有两种不同的缺损数据。首先，缺失某个样本元素的所有测量值。比如，受访人拒绝接受访问。这就是个案无应答，整个样本元素的研究变量的所有取值均缺损。另一方面，当受访人没有回答全部问题时，则是题目无应答，至少有一个样本元素的研究变量取值缺损。缺损数据可能导致偏差的估计和错误的标准误估计。

为了展示政府机构组织的大型调查的典型应答率，我们在表 4.1 中列出了本书使用的6 个真实的大规模调查（见第1 章）的应答率。这里的应答率是

指，给定抽样单位的类型，计划数据收集的数目中完成了的数据收集的份额（通常地，访谈完成或是问卷完整占整个样本规模的份额）。在 2000 年多国 PISA 的调查中，给出了国家层次应答率的中位值，因为国家间的差异较大。

表 4.1　各个调查的应答率

调查名称及介绍调查的章节	抽样单位	样本规模	应答率(%)
(1)小型芬兰健康调查，章节 5.1	个人	8 000	96
(2)职业健康保健调查，章节 5.1	单位	1 542	88
(3)PISA 2000 年调查，章节 9.4	学校	6 638	85
(4)健康保障调查，章节 9.3	家庭	6 998	84
(5)工资调查，章节 9.2	公司	1 572	80
(6)旅客交通调查，章节 9.1	个人	18 250	65

数字显示了各个调查间应答率的差异较大。应答率最高的是小型芬兰健康调查(96%)，最低的是旅客交通调查(65%)。差异产生的原因多种多样。仅仅提及几种可能性：调查主题吸引人的程度、实地调查的有效性，以及所选择的数据收集的形式。

比如，在旅客交通调查(6)中，使用了计算机辅助的电话访谈(CATI)。其中的一个问题是找出被抽中的个人的电话号码，以降低拒访。这样就将某些被抽中的个人排除在调查之外了。在两个公司调查(2)与(5)以及 2000 年 PISA 调查(3)中，使用了自己填写问卷的方式。而(1)与(4)中，受过训练的访谈人员联系受访人，并使用了传统的纸加铅笔的访谈方式来收集数据。

对应答率感兴趣的读者，可以阅读各小节中关于其所讨论的调查的简要技术报告。本书的扩展网页对这些方法有进一步的讲解。

缺损数据的类型——个案或是题目无应答——指导我们在估算过程中选择适当的校正无应答的方法。校正个案无应答的方法有各种再加权方法(reweighting methods)。题目无应答的缺损值可以使用各种推算方法来获得。接下来，给出个案无应答可能带来负面影响的例子。

个案无应答的影响

个案无应答导致样本数据的规模 $n_{(r)}$ 小于计划规模 n，因而增加标准误的估计值。这一点可以通过讨论总体总和 T 估计值 $\hat{t}_{HT}$ 的方差看出。在无放回式简单随机抽样(SRSWOR)情形下，方差为 $V_{srs}(\hat{t}_{HT}) = N^2(1-n/N)S^2/n$。其中，分母是原来的样本规模 n。当应答者数目由于无应答而减少时，分母减小，而方差增大。

个案无应答更严重的后果是，缺损的观测值将导致估算的偏差。特别是在第 k 个总体个案回答的概率 θ_k 取决于研究变量 y 的取值 Y_k 的情形下。利特尔与鲁宾(Little and Rubin, 1987)称此为不可忽略的无应答。这意味着研

究变量与应答概率间有着某种联系。比如,当回答收入问题的概率随着收入水平的增加而降低时,就产生了不可忽略的无应答。另一方面,当 Y_k 独立于 θ_k 时,无应答可以忽略。我们给出两个这样的小例子:当研究变量取值 Y_k 为常数($Y_k = \bar{Y}$)时,或是应答概率 θ_k 对于所有 k 均为常数 θ。

下面的例子是关于不可忽略无应答的。作为一个极端的例子,让我们假定在访谈调查中,某一组次级样本全部拒访。在这个例子中,总体可以分成两个次级总体,一个是应答组,另一为拒答组,其规模为 N_1 与 N_2。在实地工作之后,所有用于估算的样本数据来自于第一组,仅仅只有应答者。让总和 T 的估计值为 $\hat{t}_{HT(r)} = N \times \bar{y}_{(r)}$,其中回答数据的均值为 $\bar{y}_{(r)}$。由于所有应答者来自第一组,回答均值 $\bar{y}_{(r)}$ 的期望值等于这一组的总体均值 $\bar{Y}_1$。当总体中各组均值不等,或是 $\bar{Y}_1 \neq \bar{Y}_2$ 时,则估计值 $\hat{t}_{HT(r)}$ 是总体总和 T 的有偏估计值,有,

$$\begin{aligned} \text{BIAS}(\hat{t}_{HT(r)}) &= E(\hat{t}_{HT(r)}) - T \\ &= N\bar{Y}_1 - (N_1\bar{Y}_1 + N_2\bar{Y}_2) \\ &= N_2(\bar{Y}_1 - \bar{Y}_2) \end{aligned} \tag{4.1}$$

在实际中,得出偏差并不容易。虽然,次级总体的规模 N_2 可以大约估出,但它的均值 $\bar{Y}_2$ 几乎是未知的。同时,应当考察平均方差,而非方差。总和估计值 $\hat{t}_{HT(r)}$ 的平均方差可以写成,

$$\text{MSE}(\hat{t}_{HT(r)}) = V_{p(s)}(\hat{t}_{HT(r)}) + \text{BIAS}^2(\hat{t}_{HT(r)})。 \tag{4.2}$$

另一个麻烦是,总和估计值的方差将被低估。下面的例子演示个案无应答造成的偏差。

范例 4.1

1991 年省级人口数据个案无应答偏差。让我们假定南部自治市无法按时完成失业人数的记录。这些自治市是 Kuhmoinen, Joutsa, Luhanka, Leivonmäki,以及 Toivakka。自治市的总体可以被分成两个次级总体:应答组($N_1 = 27$)与无应答组($N_2 = 5$)。其组内总和、规模及均值为:

$T_1 = 14\ 475$　　$N_1 = 27$(应答者组别)　　$\bar{Y}_1 = 536.11$

$T_2 = 623$　　$N_2 = 5$(无应答组)　　$\bar{Y}_2 = 124.60$

$T = 15\ 098$　　$N = 32$(整个省)　　$\bar{Y} = 471.81$

当使用 SRSWOR 抽取时,选中的样本包括应答与无应答自治市。因此,根据应答组样本总和 $\hat{t}_{HT(r)}$ 得出的总和估计值的期望值为 $E(\hat{t}_{HT(r)}) = N \times \bar{Y}_1 = 32 \times 536.11 = 17\ 156$。当这一估计值被当成总体总和时,就产生了估计偏差。个案无应答导致的偏差为,

$$\begin{aligned} \text{BIAS}(\hat{t}_{HT(r)}) &= E(\hat{t}_{HT(r)}) - T \\ &= N_2(\bar{Y}_1 - \bar{Y}_2) \end{aligned}$$

$$=5\times(536.11-124.60)$$
$$=2\ 058,$$

其大小并不是可忽略的。

处理无应答的框架

在本书的第一部分,我们考察了抽样设计 $p(s)$ 引起的随机性。在无应答的例子中,我们遇到了另一种由未知的应答机制造成的随机性。给定抽样设计 $p(s)$ 情形下,规模为 n 的样本 s,它造成了应答模式 $s(r)$ 形成的未知条件概率。这让我们考虑到,在有无应答的情形下,点估计与完全基于设计的估计值不同,而相应的方差估计值含有两个部分:一个由抽样设计产生,另一由未知的应答机制产生。这个二元框架是我们猜测未知应答概率或是对其建模的前提。这一观点在伦德斯特隆与桑德尔(Lundström and Särndal, 2002)的技术报告中清晰可见。

校正无应答的两个主要方法是再加权与推算。个案无应答的校正可以使用再加权。将抽样权重 $w_k=1/\pi_k$ 用估算的应答概率 $\hat{\theta}_k$ 的倒数 $1/\hat{\theta}_k$ 来校正。由此得到分析权重或是二次权重 $w_k^*=1/(\pi_k\hat{\theta}_k)$。国家统计机构经常使用再加权来校正无应答。章节 4.1 将讨论再加权技术。

题目无应答的推算表示,观测值 y_k 的缺损值由预测值 $\hat{y}_k$ 来代替。推算的目的是为下一步的分析获得一个完整的数据矩阵。可以使用一元或多元推算。利特尔与鲁宾(Little and Rubin, 1987)从理论和实际的角度讨论了多元推算的主要思路。章节 4.2 将关注不同的推算方法。

4.1 再加权

个案无应答是指在调查数据中一定数量的抽样单位没有数据的情形。根据已知的应答者与拒访者的辅助信息,再加权可以用于应答者的观测值。作为一个简单的例子,考虑总体总和的估算。从应答者得来的数值可以乘以一个膨胀因子,以得到一个接近于最初或是计划的样本规模。一个简单的膨胀因子是应答率的倒数。比如,调查的整体应答率为 71%,恰当的膨胀因子则为 $1/0.71=1.41$。在这个无应答的模型中,假定了每一个总体元素如果被选中,有着相等的应答概率 θ。亦即,对于所有总体元素,$\theta_k=\theta, k=1,\cdots,N$,而 θ 的估计值为 $\hat{\theta}=n_{(r)}/n$。在这个非常初浅的无应答机制假设的情形下,再加权的应答概率为常数基础上的霍维茨-汤普森(HT)总体总和的估计值为,

$$\hat{t}_{HT}^{*}=\sum_{k=1}^{n_{(r)}}w_{HT,k}^{*}\times y_k=\frac{1}{\hat{\theta}}\sum_{k=1}^{n_{(r)}}w_k\times y_k=\frac{n}{n_{(r)}}\sum_{k=1}^{n_{(r)}}w_k\times y_k,\qquad(4.3)$$

其中，y_k 是应答者 k 在研究变量 y 上的观测值，$w_{HT,k}^{*}=(1/\hat{\theta})\times w_k$ 是分析权重。脚标“(r)”指应答者，所以 $n_{(r)}$ 表示样本中应答者的数目。

尽管这样的膨胀因子有时在实际中使用，但是对于应答概率的建模可以得到更好的估算。一个常用的模型是将总体分成应答同质组，简写成 RHG。这些组用 $1,\cdots,c,\cdots,C$ 来表示。其样本规模与应答者数目分别表示为 $n_1,\cdots,n_c,\cdots,n_C$ 与 $n_{1(r)},\cdots,n_{c(r)},\cdots,n_{C(r)}$。RHG 的同质性意味着，组 c 中的所有元素的应答概率相等，为 θ_c，其估计值为 $\hat{\theta}_c=n_{c(r)}/n_c$。但是，在 RHG 之间，应答概率可以不同。再加权时，使用应答概率的倒数，亦即使用估算的组应答率 $\hat{\theta}_c$。分析概率为，$w_{rhg,k}^{*}=(1/\hat{\theta}_c)\times w_k$。因而，基于 RHG 方法再加权的 HT 估计值为，

$$\hat{t}_{rhg}^{*}=\sum_{k=1}^{n_{(r)}}w_{rhg,k}^{*}\times y_k=\sum_{c=1}^{C}\left(\frac{1}{\hat{\theta}_c}\right)\sum_{k=1}^{n_{c(r)}}w_{ck}\times y_{ck}$$

$$=\sum_{c=1}^{C}\frac{n_c}{n_{c(r)}}\sum_{k=1}^{n_{c(r)}}w_{ck}\times y_{ck},\qquad(4.4)$$

其中，w_{ck}与 y_{ck}分别是抽样权重与组 c 中应答者 k 在 y 的取值。

这一个案无应答的校正比前一个方法更有效。因为，它在用模型估算应答概率时，更为有效地使用了无应答的结构的信息。当已知辅助变量 z 的取值 z_k，并且 z 与研究变量 y 相关时，可以使用再加权比率估计值。其权重为，$w_{rat,k}^{*}=[(1/\hat{\theta})\times(\bar{z}/\bar{z}_{(r)})]\times w_k$。其中，$\bar{z}$ 为辅助变量 z 在所用选中个案的均值，而 $\bar{z}_{(r)}$ 则是应答个案的均值，$\hat{\theta}=n_{(r)}/n$。相应的，基于比率方法再加权的 HT 估计值为，

$$\hat{t}_{rat}^{*}=\sum_{k=1}^{n_{(r)}}w_{rat,k}^{*}\times y_k=\frac{\bar{z}}{\hat{\theta}\times\bar{z}_{(r)}}\sum_{k=1}^{n_{(r)}}w_k\times y_k=\frac{n\times\bar{z}}{n_{(r)}\times\bar{z}_{(r)}}\sum_{k=1}^{n_{(r)}}w_k\times y_k。\qquad(4.5)$$

接下来，我们转向总和的再加权 HT 估计值的方差估算。在基于设计的推论情形中，抽样权重为已知常数 $w_k=1/\pi_k$。再加权中，这些常数乘上一个因具体再加权方法与样本而定的加权因子。这就导致了另外一个需要测量与纳入总和估计值基于设计的方差成分。我们用 V_{rew} 来表示这一成分，“rew”（reweighting——译者注）是指再加权。方差成分可以用上面定义的处理个案无应答的框架来估算。为了这个目的，我们在概念上将样本选取分成两个阶段：根据抽样设计 $p_{(s)}$ 选取样本 s，以及从选取的样本 s 中获得应答数据 $s_{(r)}$。这样的抽样分割使得我们可以分别估算第一和第二阶段的方差。写成 V_{sam} 的第一个部分表示抽样设计引起的方差，写成 V_{rew} 的第二个部分表示由未知应答机制引起的方差。如桑德尔（Särndal，1996）所假定这两个部分相互独立，则总和 T 的再加权 HT 估计值 $\hat{t}_{HT}^{*}$的方差可以被分解成，

$$V(\hat{t}_{HT}^{*})=V_{sam}(\hat{t}_{HT}^{*})+V_{rew}(\hat{t}_{HT}^{*}),\qquad(4.6)$$

其中，$V_{sam}(\hat{t}^*_{HT})$是基于应答数据的基本 HT 估计值 $\hat{t}_{HT}$的设计方差，而 $V_{rew}(\hat{t}^*_{HT})$是由再加权方法引起的方差部分。在范例 4.2 中，将计算三个再加权估计值 $\hat{t}^*_{HT}$，$\hat{t}^*_{rhg}$，$\hat{t}^*_{rat}$及其方差的各个部分。

范例 4.2

在 1991 年省级人口数据的 SRSWOR 样本中，使用再加权来校正个案无应答。表 4.2 给出了数据。让我们假定两个个案无应答，Kuhmoinen 与 Toivakka。注意，辅助变量 HOU85 的取值对于无应答个案也是已知。最初的样本规模为 8 个自治市。估计的应答率为，$\hat{\theta} = n_{(r)}/n = 6/8 = 0.75$。同时，3 个选中的自治市为城镇（应答同质组 $c = 1$），其余 5 个为农村自治市（应答同质组 $c = 2$）。因为所有城镇自治市都回答了，相应的估计的应答概率第一组为 $\hat{\theta}_1 = 3/3 = 1.00$，第二组为 $\hat{\theta}_2 = 3/5 = 0.60$。整个样本（$n = 8$）的辅助变量 HOU85 的均值为 $\bar{z} =$ 5 154.75，而应答样本（$n = 6$）中的 HOU85 的均值为 $\bar{z} =$ 6 490.17。有了这些背景信息，我们可以计算前面引入的变量 UE91 总和的再加权估计值 $\hat{t}^*_{HT}$，$\hat{t}^*_{rhg}$及 $\hat{t}^*_{rat}$。

表 4.2　含两个无应答的 1991 年省级人口数据的一个简单随机样本，应答同质组与单位无应答修正的权重

抽样设计标识			元素标签	回应变量数据		应答同质组	无应答模型再加权		
							REW_HT	RHG	RATIO
STR	CLU	WGHT		UE91	HOU85		$w^*_{HT,k}$	$w^*_{rhg,k}$	$w^*_{rat,k}$
1	18	4	Kuhmoinen	··	1 463	2	··	··	··
1	30	4	Toivakka	··	834	2	··	··	··
1	26	4	Pihtipudas	331	1 946	2	5.333 3	6.666 7	4.235 9
1	31	4	Uurainen	219	932	2	5.333 3	6.666 7	4.235 9
1	15	4	Konginkangas	142	556	2	5.333 3	6.666 7	4.235 9
1	1	4	Jyväskylä	4 123	26 881	1	5.333 3	4.000 0	4.235 9
1	4	4	Keuruu	760	4 896	1	5.333 3	4.000 0	4.235 9
1	5	4	Saarijärvi	721	3 730	1	5.333 3	4.000 0	4.235 9

缺损值用"··"表示。

为了计算二次权重，我们需要首先定义恰当的应答同质组。在这个例子中，估计值 $\hat{t}^*_{HT}$与 $\hat{t}^*_{rat}$的自然分组是整个样本，而估计值 $\hat{t}^*_{rhg}$则根据是否为城镇分成两个应答同质组。对于估计值 $\hat{t}^*_{HT}$，我们使用简单的再加权方法：应答者的二次权重为 $w^*_{HT,k} = (1/\hat{\theta}) \times w_k = (1/0.75) \times 4 = 5.333\ 3$。对于估计值 $\hat{t}^*_{rhg}$，第一个应答同质组（城镇）的二次权重为 $w^*_{rhg,1} = (1/\hat{\theta}_1) \times w_k = (1/1) \times 4 = 4$。由于所有样本都回答了，它等于抽样权重。而第二个应答同质组（农村）中，

$w^*_{rhg,2} = (1/\hat{\theta}_2) \times w_k = (1/0.60) \times 4 = 6.6667$。对于比率估计值，整个样本则被当成应答同质组。我们使用在计算校正权重的例子中给出的公式（见章节3.3的比率估算）。但在这里，使用样本中估算得来的辅助变量的总体均值（或总和）。应答者的二次权重为 $w^*_{rat,k} = (1/\hat{\theta}) \times (\bar{z}/\bar{z}_{(r)}) \times w_k = [(n \times \bar{z}) / (n_{(r)} \times \bar{z}_{(r)})] \times w_k$。对于 SRSWOR 而言，各个应答者的二次权重相等。选中样本中应答者的实际数值为 $w^*_{rat} = [(8 \times 5\,154.75) / (6 \times 6\,490.17)] \times 4 = 4.2359$。

使用计算出的二次权重，可以得出点估计值及其方差。UE91 总和 T 的点估计就是从应答者数据中计算出来的再加权 HT 估计值。表4.3给出了估计值。我们关心方差估算，因为它现在包含两个部分：抽样设计引起的方差估计值 $\hat{v}_{sam}$ 以及应答机制引起的方差估计值 $\hat{v}_{rew}$。我们假定无应答在各个应答同质组内部可以忽略。

表4.3 各种再加权方法下的总和及其方差估计值

方法与估算值	总和估计值	$\hat{v}(\hat{t})$	$\hat{v}_{sam}$	$\hat{v}_{rew}$
应答者数据（$n_{(r)}=6$）$\hat{t}_{HT(r)}$	33 579	$17\,988^2$	$17\,988^2$	0
再加权估计值 $\hat{t}^*_{HT}$	33 579	$17\,988^2$	$14\,967^2$	$9\,978^2$
应答同质组 $\hat{t}^*_{rhg}$	27 029	$14\,983^2$	$14\,967^2$	694^2
比例估计值 $\hat{t}^*_{rat}$	26 669	$14\,988^2$	$14\,967^2$	786^2
"完全应答"（$n=8$）$\hat{t}_{HT}$	26 440	$13\,282^2$	$13\,282^2$	0

数据采自表4.2中的1991年省级人口的简单随机样本。

由于抽样设计为 SRSWOR，我们使用相应的总和的设计方差，

$$V_{sam}(\hat{t}^*_{HT}) = N^2\left(1 - \frac{n}{N}\right) \times S^2_{(r)}/n_{(r)} \tag{4.7}$$

其中，$S^2_{(r)} = \sum_{k=1}^{N_{(r)}} (Y_k - \bar{Y}_{(r)})^2/(N_{(r)} - 1)$ 是从总体 U 中的应答部分 $U_{(r)}$ 计算得来。

这个例子中这一部分的估计值为，

$$\begin{aligned}\hat{v}_{sam}(\hat{t}^*_{HT}) &= N^2\left(1 - \frac{n}{N}\right) \times \hat{s}^2_{(r)}/n_{(r)} \\ &= 32^2\left(1 - \frac{8}{32}\right) \times 1\,527.59^2/6 \\ &= 14\,967^2,\end{aligned}$$

其中，$\hat{s}^2_{(r)} = \sum_{k=1}^{n_{(r)}} (y_k - \bar{y}_{(r)})^2/(n_{(r)} - 1)$，并且从应答数据中估算得来。

这一部分方差对于各个再加权估计值是相同的。整个方差的再加权部分取决于使用的再加权方法。接下来，估算各个再加权方法中的 V_{rew}。注意，当 HT 估计值 $\hat{t}_{HT(r)}$ 从应答者数据中估算时，并不含有再加权引起的方差部分。

1. 再加权估计值 $\hat{t}_{HT}^*$。在第一个再加权 HT 估计值的例子中，方差 $V_{rew}(\hat{t}_{HT}^*)$ 为，

$$V_{rew}(\hat{t}_{HT}^*) = N^2\left(1 - \frac{n_{(r)}}{n}\right) \times S_{(r)}^2 / n_{(r)}, \quad (4.8)$$

其中，$S_{(r)}^2 = \sum_{k=1}^{N_{(r)}} (Y_k - \overline{Y}_{(r)})^2 / (N_{(r)} - 1)$ 是从总体 U 中的应答部分 $U_{(r)}$ 计算得来。这一部分的估计值为，

$$\hat{v}_{rew}(\hat{t}_{HT}^*) = N^2\left(1 - \frac{n_{(r)}}{n}\right) \times \hat{s}_{(r)}^2 / n_{(r)}$$

$$= 32^2\left(1 - \frac{6}{8}\right) \times 1\ 527.59^2/6 = 9\ 978.18^2,$$

其中，$\hat{s}_{(r)}^2$ 从应答数据中估算得来。

2. 应答同质组估计值 $\hat{t}_{rhg}^*$。我们有两个 RHG，其样本规模为 $n_1 = 3$ 与 $n_2 = 5$。从这些数字，可以估算出相应的次级总体规模，它们是，

第一组次级总体：$\hat{N}_1 = (n_1/n) \times N = (3/8) \times 32 = 12$，

第二组次级总体：$\hat{N}_2 = (n_2/n) \times N = (5/8) \times 32 = 20$。

应答同质组估计值 $\hat{t}_{rhg}^*$ 方差的再加权部分为，

$$V_{rew}(\hat{t}_{rhg}^*) = \sum_{c=1}^{C} \hat{N}_c^2\left(1 - \frac{n_{c(r)}}{n_c}\right) \times S_{c(r)}^2 / n_{c(r)}, \quad (4.9)$$

其中，$S_{c(r)}^2 = \sum_{k=1}^{N_{c(r)}} (Y_{ck} - \overline{Y}_{c(r)})^2 / (N_{c(r)} - 1)$ 从各个应答同质组分别计算得出。次级总体 U_c 中的应答数目用 $N_{c(r)}$ 来表示，$c = 1, 2$。相应的估计值 $\hat{v}_{rew}(\hat{t}_{rhg})$ 用公式 4.9 及应答者数据计算得出，其中 $S_{c(r)}^2$ 由其估计值 $\hat{s}_{c(r)}^2$ 替代。因此，我们有，

$$\hat{v}_{rew}(\hat{t}_{rhg}^*) = 12^2\left(1 - \frac{3}{3}\right) \times 1\ 952.99^2/3 + 20^2\left(1 - \frac{3}{5}\right) \times 95.04^2/3$$

$$= 0 + 694.07^2$$

$$= 694.07^2。$$

3. 再加权比率估计值 $\hat{t}_{rat}^*$。首先，我们推出残差变量 $E_{k(r)} = Y_{k(r)} - (\overline{Y}_{(r)} / \overline{Z}_{(r)}) \times Z_{k(r)}$。注意，这一残差从总体中应答部分计算出来。估计值 $\hat{t}_{rat}^*$ 方差的再加权部分为，

$$V_{rew}(\hat{t}_{rat}^*) = N^2\left(1 - \frac{n_{(r)}}{n}\right) \times S_{E(r)}^2 / n_{(r)}, \quad (4.10)$$

其中，$S_{E(r)}^2 = \sum_{k=1}^{N_{(r)}} (E_{k(r)} - \overline{E})^2 / (N_{(r)} - 1)$ 及 $\overline{E} = \sum_{k=1}^{N_{(r)}} E_{k(r)} / N_{(r)}$。

残差 $E_{k(r)}$ 从应答者数据中估算出来，$\hat{e}_{k(r)} = y_{k(r)} - (\overline{y}_{(r)} / \overline{z}_{(r)}) \times z_{k(r)}$。

在这个特别的例子中，方差 $\hat{v}_{rew}(\hat{t}_{rat}^*)$ 的再加权部分为，

$$\hat{v}_{rew}(\hat{t}^*_{rat}) = N^2\left(1 - \frac{n_{(r)}}{n}\right)\hat{s}^2_{\hat{e}_{(r)}}/n_{(r)} = 32^2\left(1 - \frac{6}{8}\right) \times 120.29^2/6 = 785.73^2,$$

其中，$\hat{s}^2_{\hat{e}(r)} = \sum_{k=1}^{n_{(r)}} (\hat{e}_{k(r)} - \bar{\hat{e}}_{(r)})^2/(n_{(r)} - 1)$从应答者数据中计算得出。

抽样比例定义为，样本中的应答者数目除以目标总体中应答同质组的估计或是实际规模。估计值$\hat{t}^*_{HT}$与$\hat{t}^*_{rat}$以整个样本作为应答同质组。因此，抽样比例为$n_{(r)}/N = 6/32 = 0.1875$。对于估计值$\hat{t}^*_{rhg}$，第一个应答同质组的抽样比例为$n_{1(r)}/\hat{N}_1 = 3/12 = 0.25$，第二个应答同质组的抽样比例为$n_{2(r)}/\hat{N}_2 = 3/20 = 0.15$。

表4.3列出了计算结果。除了再加权估计值以外，也加入了两个参照估计值。估计值$\hat{t}_{HT(r)} = N \times \bar{y}_{(r)}$从应答者数据中计算得出。在这个例子中，抽样比例为$n_{(r)}/N = 6/32 = 0.1875$。为了达到公正的比较，基本的基于设计的估计值$\hat{t}_{HT}$由最后一行"完整应答"中的数字计算得出。这里的抽样比例是$n/N = 8/32 = 0.25$。由式4.7分别估算应答者数据（$n = n_{(r)} = 6$）与完整应答（$n = 8$）的方差部分$\hat{v}_{sam}$。最后一列以$\hat{v}_{rew}$为列标，显示对于应答者数据（第一行）与完整应答（末行）没有因再加权引起的方差部分。

再加权估计值的令人满意的特征是，它产生尽可能接近完整应答的数值。在这一点上，应答者数据估计值$\hat{t}_{HT(r)}$与再加权HT估计值$\hat{t}^*_{HT}$均给出较差的结果。点估计$\hat{t}_{HT(r)} = \hat{t}^*_{HT} = 33\ 579$与"完整应答"估计值$\hat{t}_{HT} = 26\ 440$相去甚远。方差估计值也是如此，$\hat{v}(\hat{t}_{HT(r)}) = \hat{v}(\hat{t}^*_{HT}) = 17\ 988^2 > 13\ 282^2$。再加权HT估计值产生较差结果的原因是，简单的应答机制假定了所有总体元素有着同一常数应答概率$\hat{\theta}$。应答同质组估计值$\hat{t}^*_{rhg}$与比率估计值$\hat{t}^*_{rat}$使用了从样本数据中搜集的辅助信息。这些估计值的使用是建立在更恰当的模型假设之上的。正如这里的例子，当这些假设符合实际时，这两个估计值与"完整应答"估计值十分接近。

4.2 推　算

题目无应答是指，对于某一样本元素，分析数据中含有某些数值，但至少缺损一个题目的数值。当使用调查估算的计算机程序来运算这样的数据矩阵时，分析中任一变量缺损其中一个数值的个案均将被排除在外。另外，某些分析需要完整的数据矩阵。这将导致丢失掉其他数据并不缺损的变量的信息。所以，要努力获得更完整的数据。为了达到这一目的，发展出了各种推算技术。

推算是指，数据矩阵中样本元素 k 在研究变量 y 的缺损值由推算值 $\hat{y}_k$ 所替代。比如，某些计算机软件含有均值推算的技术。它使用应答数据计算出的研究变量的整体应答均值 $\bar{y}_{(r)}$ 来填补该变量的缺损值。这里，元素 k 的推算值 $\hat{y}_k = \bar{y}_{(r)}$。但是，范例 4.3 将展示这一方法的不利方面。更高级的方法是使用总体框架或是原有样本中的辅助信息，更加符合实际地以模型来模拟缺损值。

均值推算并不使用样本数据中可能存在的辅助信息。与前面一样，与研究变量 y 相关并且对于所有样本元素已知的辅助变量 z，可以用于推算方法中。比如，我们可以使用辅助变量 z 的样本取值来生成一个两个样本元素间的距离 $|z_l - z_k|$，这里 $l \neq k$。距离最小的样本元素被称作最近值。当元素 k 属于无应答组，元素 l 属于应答组，我们用数值 y_l 来替代元素 k 的缺损值，得到推算值 $\hat{y}_k = y_l$。因此，样本元素 l 是元素 k 的供值者。注意，这个估计值是实际观测到的。与再加权的情形一样，这里也可以使用变量估算。现在，我们使用等式 $\hat{y}_k = z_k \times (\bar{y}_{(r)}/\bar{z}_{(r)})$ 来预测各个缺损值，其中，$\bar{y}_{(r)}$ 与 $\bar{z}_{(r)}$ 分别是研究变量与辅助变量的应答均值。比如，不完整的数据可以用热板推算。在热板推算中，从样本数据中随机选取一个测量值，并用它来填补缺损值。所有这些方法均是单次推算方法。

单一的缺损值可以由多次推算方法中的两个或多个推算值来填补。对每一个缺损值，这需要独立地完成。当我们对每一个缺损值重复 m 次后，我们得到 m 个可用于统计分析的完整数据。可以使用原有的从样本设计 $p(s)$ 中推出的加权方法。我们应当在估计值方差公式的基础上，评估并纳入推算方差的部分。在单次推算的情形下，预测值 $\hat{y}_k$ 代替了缺损值，可以使用公式 4.6，并用由推算引起的方差部分 V_{imp} 代替 V_{rew}，有

$$V(\hat{t}^*) = V_{sam} + V_{imp}。 \tag{4.11}$$

在多次推算中，对每一缺损值，我们预测 m 个数值 $\hat{y}_1, \cdots, \hat{y}_j, \cdots, \hat{y}_m$。我们有 m 个“完整”的数据。为了合并这些结果，我们首先定义感兴趣的参数的多次推算估计值。比如，总和的估计值为，

$$\hat{t}^*_{mi} = \frac{1}{m} \times \sum_{j=1}^{m} \hat{t}^*_j, \tag{4.12}$$

其中，$\hat{t}^*_j$ 是总和的一个估计值，$\hat{v}_{p(s)}(\hat{t}^*_j)$ 是第 j 个“完整”数据的方差估计值，$j=1,\cdots,m$。$\hat{t}^*_{mi}$ 的方差估计值包含两个部分：推算内的方差部分与推算间的方差部分。推算内的方差由 m 个方差估计值 $\hat{v}_{p(s)}(\hat{t}^*_j)$ 的均值得出，表示方差 V_{sam}。推算间的方差与 $\hat{t}^*_j$ 的变动有关联。这一部分可以被看成推算引起的方差 V_{imp}。在多次推算的情形下，总和的方差估计值为，

$$\hat{v}(\hat{t}^*_{mi}) = \hat{v}_{sam} + \hat{v}_{imp}$$

$$= \left[\frac{1}{m} \times \sum_{j=1}^{m} \hat{v}_{p(s)}(\hat{t}_j^*) \right] + \left[\left(1 + \frac{1}{m}\right) \times \sum_{j=1}^{m} \frac{(\hat{t}_j^* - \hat{t}_{mi}^*)^2}{m-1} \right] \quad (4.13)$$

其中，$\hat{v}_{p(s)}(\hat{t}_j^*)$是从样本设计 $p(s)$ 情形下的第 j 个完整数据中计算出来的方差估计值，$(1+(1/m))$是对有限 m 的校正。

在实际中，m 通常是一个小数目，至少为 $m=2$，但最好为 3 到 5。范例 4.3 演示不同推算方法的方差的估算。

范例 4.3

我们对从 1991 年省级人口数据中以 SRSWOR 抽取的样本推算两个缺损值。使用与范例 4.2 相同的样本，含有 $n=8$ 个自治市。两个自治市（Kuhmoinen 与 Toivakka）中的研究变量 UE91 含有缺损值。表 4.4 中的数据用"··"来表示缺损值。变量 HOU85 是没有缺损值的辅助变量。

表 4.4 单一推算方法得出的完整数据（1991 年省级人口）

序号 k	元素标签	回应变量数据		模型得出的推算数据			完全回应
		UE91	HOU85	应答均值	最近值	比率估算	
18	Kuhmoinen	..	1 463	1 049.33*	331*	236.54*	187
30	Toivakka	..	834	1 049.33*	219*	134.84*	127
1	Jyväskylä	4 123	26 881	4 123	4 123	4 123	4 123
4	Keuruu	760	4 896	760	760	760	760
5	Saarijärvi	721	3 730	721	721	721	721
15	Konginkangas	142	556	142	142	142	142
26	Pihtipudas	331	1 946	331	331	331	331
31	Uurainen	219	932	219	219	219	219

"*"：推算值；".."：缺损值。

我们使用 4 种推算方法来填补样本数据。第一种是应答者均值推算方法。应答者（$n_{(r)}=6$）的均值为 $\bar{y}_{(r)}=1\ 049.33$。用这个整体上的均值来代替两个缺损值。第二种和第三种无应答模型将变量 HOU85 用作辅助变量 z。第二种方法叫做最近值推算。对于无应答元素 k，我们选择潜在的供值者（潜在的供值者是属于变量 y 的应答组的样本元素）中距离 $|z_k - z_l|$ 最小的应答元素 l 的取值。按最小距离的标准，Pihtipudas（$y_{26}=331$）是 Kuhmoinen 的供值者（$|1\ 949-1\ 463|=486$），而 Uurainen（$y_{31}=219$）是 Toivakka 的供值者（$|932-834|=98$）。在第三种模型中，我们使用比率估算。我们从应答数据中计算比率 $\hat{B}=\bar{y}_{(r)}/\bar{z}_{(r)}=1\ 049.33/6\ 490.17=0.161\ 7$，得出 $\hat{y}_k=\hat{B}\times z_k$，即 Kuhmoinen 的 $\hat{y}_{18}=0.161\ 7\times 1\ 463=236.57$，Toivakka 的 $\hat{y}_{30}=0.161\ 7\times 834=134.86$。表4.4中有填补后的样本数据。注意，均值推算与比率推算使用预测值来代替缺损观测值。正是由于此，这些数值保留两位小数。另一方面，以最

近值为供值者的推算给出整数值。多次推算也是如此，因为我们使用任一应答数值均为潜在供值者的热板推算。

每个单次推算得出一个完整的数据，共产生 3 个完整数据供估算所用。由于我们的抽样设计为 SRSWOR，我们使用的抽样权重为常数，$w_k = 4$。但是，方差估计值中有了新的情况，现在它包含两个部分（见公式 4.11）。在总和估算中，抽样方差的估计值为，

$$\hat{v}_{sam}(\hat{t}^*_{HT}) = N^2\left(1 - \frac{n}{N}\right) \times \hat{s}^2_{(r)}/n_{(r)}$$

$$= 32^2\left(1 - \frac{8}{32}\right) \times 1\,527.59^2/6 = 14\,967^2$$

其中，$\hat{s}^2_{n(r)} = \sum_{k=1}^{n_{(r)}} (y_k - \bar{y}_{(r)})^2/(n_{(r)} - 1)$ 从应答数据中计算得出。这一方差部分对于各种推算方法相同。所有单次推算方法的推算方差部分 V_{imp} 的估计值为，

$$\hat{v}_{imp}(\hat{t}^*_{HT}) = N^2\left(1 - \frac{n_{(r)}}{n}\right) \times \frac{\sum_{k=1}^{n_{(r)}} (\hat{e}_k - \bar{\hat{e}})^2}{n_{(r)} - 1} / n_{(r)} \qquad (4.14)$$

其中，$\bar{\hat{e}} = \sum_{k=1}^{n_{(r)}} \hat{e}_k/n_{(r)}$ 是残差 $\hat{e}_k = y_k - \hat{y}_k$ 的均值。在均值推算中，残差为 $\hat{e}_k = y_k - \bar{y}_k$。使用最近值为供值者，得出残差 $\hat{e}_k = y_k - y_{k(l)}$，其中 $y_{k(l)}$ 为供值者 l 在 y 的取值。比率估算导致的残差为 $\hat{e}_k = y_k - (\bar{y}_{(r)}/\bar{z}_{(r)}) \times z_k$。将这些变量带入式 4.14中，我们得到估算的推算方差部分为，

$$\hat{v}_{imp}(\hat{t}^*_{rm}) = 32^2\left(1 - \frac{6}{8}\right) \times 1\,527.59^2/6 = 9\,978.18^2$$

$$\hat{v}_{imp}(\hat{t}^*_{nn}) = 32^2\left(1 - \frac{6}{8}\right) \times 1\,365.62^2/6 = 8\,920.20^2$$

$$\hat{v}_{imp}(\hat{t}^*_{ra}) = 32^2\left(1 - \frac{6}{8}\right) \times 120.29^2/6 = 785.73^2。$$

注意，比率模型的推算方差最小。

接下来，我们转到多次推算方法。在这个简单的练习中，我们使用 5 次重复的热板推算。注意，使用热板推算仅仅是为了在这个限制颇严的小规模数据上演示多次推算的基本原则。为了实用的目的，发展和计算机化了更加复杂的多次推算技术。有很多关于其他技术的文献。读者可以参阅谢弗（Schafer, 2000）的书以及鲁宾（Rubin, 1996）的文章。

每一次运算，缺损值由应答数据中随机选取的数值所替代。这样的过程产生了 5 个完整数据，表 4.5 列出了这些数据。失业人数总和的点估计 $\hat{t}^*_{mi}$ 是 5 个单独数据的、同一总和估计值 $\hat{t}^*_j$ 的均值。因此，由式 4.12，我们有，

$$\hat{t}^*_{mi} = (1/5)(28\,792 + 31\,108 + 28\,944 + 44\,716 + 29\,100) = 32\,532。$$

由(4.13),估计值$\hat{t}_{mi}^*$的方差分解为推算内的部分与推算间的部分。推算内差异的元素是估计值$\hat{t}_j^*$的设计方差估计值的5个数据估计值。因此,式4.13的第一项为,

$$\hat{v}_{sam}=\frac{1}{m}\times\sum_{j=1}^{m}\hat{v}_{p(s)}(\hat{t}_j^*)=\frac{1}{5}\times\left(1-\frac{8}{32}\right)\times 32^2\times$$
$$(1\ 330.715^2+1\ 298.982^2+1\ 325.416^2+$$
$$1\ 699.989^2+1\ 324.716^2)/8=13\ 758.87^2$$

其中,$\hat{v}_{p(s)}(\hat{t}_j^*)=\hat{v}_{srswor}(\hat{t}_j^*)$或是一个抽样设计为SRSWOR的总和的方差估计值。

表4.5 多次推算($m=5$)得到的推算数据。各个完整数据使用了热板推算(1991年省级人口)

序号	元 素	回应变量数据	带"*"推算值的重复样本					完全回应
		UE91	1	2	3	4	5	
18	Kuhmoinen	..	760*	760*	721*	4 123*	760*	187
30	Toivakka	..	142*	721*	219*	760*	219*	127
1	Jyväskvlä	4 123	4 123	4 123	4 123	4 123	4 123	4 123
4	Keuruu	760	760	760	760	760	760	760
5	Saarijärvi	721	721	721	721	721	721	721
15	Konginkangas	142	142	142	142	142	142	142
26	Pihtipudas	331	331	331	331	331	331	331
31	Uurainen	219	219	219	219	219	219	219
	Mean	1 049.33	899.75	972.12	904.50	1 397.38	909.37	826.25
	STD(y)	1 527.59	1 330.71	1 298.98	1 325.42	1 699.99	1 324.72	1 355.15

"*":推算值;"..":缺损值。

相应的差异间或是推算间的方差的估算为,

$$\hat{v}_{imp}=\left(1+\frac{1}{m}\right)\times\sum_{j=1}^{m}\frac{(\hat{t}_j^*-\hat{t}_{mi}^*)^2}{m-1}$$
$$=1.2\times 6\ 876.444^2=7\ 532.39^2$$

将这两项相加,得到估计值$\hat{t}_{mi}^*$的方差估计值,

$$\hat{v}(\hat{t}_{mi}^*)=\hat{v}_{sam}+\hat{v}_{imp}=13\ 758.87^2+7\ 532.39^2=15\ 686.86^2。$$

表4.6给出各种推算方法的结果。同样的,为了比较,末行表示"完整应答"情形下的估计值及其方差。这一行作为参照。如果推算方法较好,它产生的数值应当接近于"完整应答"估计值。点估计应当如此(但方差估计值例外,因为它包含另外的推算方差项)。

应答均值推算给出了与"无修正"方法相同的总和估计值(33 579),但是低估了方差,除非加上推算方差($\hat{v}_{imp}=9\ 978^2$)。更加高级的无应答方法——"最近值"与"比率估算"——得出了更接近于从"完整应答"数据计算出的参

照值。"最近值方法"得出 $\hat{t}_{nn}^{*}=27\ 384$。比率模型使用辅助信息,其估计值为 $\hat{t}_{ra}^{*}=26\ 669$,很接近于参照值 $\hat{t}_{HT}=26\ 640$。尽管,推算导致的方差补偿仅仅是较小的增加,推算方差为 $\hat{v}_{imp}=786^2$。

表 4.6　各种方法下总和及其标准误估计值(1991 年省级人口)

模型类型	估算量	总和估计值	$\hat{v}(\hat{t}_*)$	$\hat{v}_{sam}$	$\hat{v}_{imp}$
无修正($n_{(r)}=6$)	$\hat{t}_{HT(r)}$	33 579	$17\ 988^2$	$17\ 988^2$	0
应答者均值	$\hat{t}_{ma}^{*}$	33 579	$17\ 988^2$	$14\ 967^2$	$9\ 978^2$
多次推算($m=5$)	$\hat{t}_{mi}^{*}$	32 532	$15\ 686^2$	$13\ 752^2$	$7\ 532^2$
最近值	$\hat{t}_{nn}^{*}$	27 384	$17\ 424^2$	$14\ 967^2$	$8\ 920^2$
比率估算	$\hat{t}_{ra}^{*}$	26 669	$14\ 988^2$	$14\ 967^2$	786^2
完全回应($n=8$)	$\hat{t}_{HT}$	26 440	$13\ 282^2$	$13\ 282^2$	0

多次推算有所不同。点估计 $\hat{t}_{mi}^{*}=32\ 532$ 显然大于"完整应答"数据的估计值。另一方面,由公式 4.13 计算得出的整体上的方差为 $\hat{v}(\hat{t}_{mi}^{*})=15\ 686^2$,小于最近值推算与应答均值推算的方差。

推算有两方面不同的影响。第一,可以推算缺损值的替代数值;第二,推算影响我们感兴趣的估计值的标准误。推算的明显优势在于分析者有了一个完整的数据矩阵。但是,当推算给出带有偏差的数值,分析的结果可能会误导。所有的都取决于推算模型在多大程度上成功地贴近无应答。当应答同质组内的无应答可以忽略时,应答均值是应答同质组——包括缺损值在内的所有同组元素——的无偏估计值。但是,由于推算值是估计值,它们自己的方差部分应当纳入基本估计值的方差之中。

4.3　本章小结与更多的文献

这一小节的目的是,给出调查结果的更宽泛的概览。这样的调查结果仅仅考虑了基于设计估计值及其随机抽样中的随机性产生抽样误差。通过简要地考察调查误差的不同来源,我们讨论了非抽样误差。这些不同的误差是整体调查误差的组成部分。整体调查误差的概念难以定义,在实际中就更难测量了。一个原因是因为不同的调查误差之间并不相互独立。但是,为了实际目的,可以分开考虑不同类别的调查误差,并逐一寻找降低它们的策略。因此,整体调查误差可以降低。

大规模抽样调查面临无应答的情况导致不完整的数据。因为绝大多数数据分析的计算机软件假定完整数据,所以作为数据搜集后的第一个步骤是,清理和修正缺损值。无应答包括个案无应答或者题目无应答,它造成含有偏差

的估算以及错误的标准误估计值。有效的方法在降低无应答的数据校正过程中相当重要。

有多种方法来校正无应答。我们介绍了使用样本数据中的辅助信息来对无应答建模的实用校正方法。辅助数据可以从普查或是商业登记中获得。这些方法的区别在于,在多大程度上依赖辅助信息。格罗夫斯等(Groves et al., 2001)讨论了社会调查中的无应答,迪尔曼(Dillman, 1999)则讨论了商业调查中的无应答。

在大规模调查中经常会遇到框架涵盖误差、数据处理误差以及测量误差,美国统计方法联邦委员会(Federal Committee on Statistical Methodology, 2001)在一篇出版的政策文章中讨论了这些。以下使用这篇文章中的定义。

框架涵盖误差是指没有将一些目标总体元素纳入抽样的框架之中(涵盖不足)以及将一些不应进入抽样框架的对象纳入了其中(过度涵盖)所造成的偏差。涵盖误差的来源是抽样框架本身。因此,评估抽样框架的质量信息以及目标总体的完备情况很重要。涵盖误差的测量方法依赖于调查之外的方法:比如,将调查估计值与其他不相关联的结果相比较,或是将两个登记表逐一对照。涵盖误差并不留下明显的痕迹,它们只能通过与其他资料相比较得出。常用的方法是加总比较与逐一对照。可以将研究总体中的年龄、性别和其他人口特征与普查登记相比较。测量涵盖误差的第二个方法基于逐一比较。这一方法假定已有另外的总体名单,或是通过普查、调查、记录可以建立这样的名单。两个名单上均没有的总体是无法观测的。但是,当两个名单相互独立时,可以估计涵盖误差。在 CATI 中常用的测量方法是用名义上的样本规模除以样本中可以找出电话号码的数目。章节 9.1 将给出一个例子。

数据处理误差发生在调查实际收集上来之后,通常是将数据转换成机器阅读形式以便统计分析的过程当中。数据处理误差包括数据录入、编码以及校正误差,因而也包括损坏的数据记录。误差率由质量控制决定,但是,近年来研究者倾向于连续管理的实践方式。比如,校正是发现和修正数据收集所引起的数据记录上的错误的程序。校正的原则,或是校正本身被用于发现缺损、错误或是可疑的数值。总体而言,这一过程由计算机辅助方法来完成。比如,考克斯等(Cox et al., 1995)与库珀等(Couper et al., 1998)使用了大量章节来讨论发现和处理商业与社会调查中数据错误的方法。发布官方统计数字的国家统计机构也发展出了监测与修正数据处理误差的自动程序。

测量误差是指变量观测值与该变量的真实但无法观测的数值间的差。数据收集中的测量误差有 4 个主要来源:问卷(呈现和索要信息的正式形式)、访谈员对于问题回答的影响(访谈员效应)、数据收集的方式,以及应答者(信息索取的对象)。这些方面组成了数据收集,每一个方面都可能在测量阶段引入误差。比如,测量误差可能产生于应答者对问题的回答,包括误解问题的

意思、无法准确回忆起信息，以及无法正确组织其回答（比如，无法正确加总某些数值）。测量误差难以数量化，通常需要特别的、昂贵的研究。回访项目、记录检查、行为编码、认知测验、随机实验是少数用于数量化测量误差的方法。测量误差的一个例子是访谈员效应，可以用章节 2.3 介绍的组内相关系数 ρ_{int} 来测量。比如，比莫尔等（Biemer et al., 1991）以强烈的实用背景讨论了商业和社会调查中的测量误差。

在这一背景中，整体调查质量是一个有趣的概念。它指包括调查误差的抽样和各个非抽样部分的多层面特征。格罗夫斯（Groves, 1989）从不同的角度讨论了这一概念的双重性——调查误差可分为观测性误差与非观测性误差。同时，他分别分析了不同误差的效应及其它们怎样影响估计值的偏差与方差。在分析几个实际调查例子之后，得出的结论是，调查质量确实是多层面的特征，并且不同的部分可能相互关联。所以，商业和社会调查质量概览以一套精心定义的指标来建构与交流，而非整体调查误差的一个数值。普拉特克与桑德尔（Platek and Särndal, 2001）强烈赞成这一观点。

第5章 线性化与方差估算中样本的再使用

Linearization and Sample Reuse in Variance Estimation

在本章与第 7 章、第 8 章中，我们讨论常见于社会、健康及教育科学的复杂分析性调查中的估算、检验和建模方法。在分析性调查中，方差估算需要样本均值与样本在整个总体中——更重要的是子总体中——的比例的标准误。在建模的过程中，为了获得合适的检验统计量，需要估算的模型系数——诸如回归系数——的方差估计值。在章节 5.2 中，子总体的均值与比例被定义为比率估计值。这些非线性估计值方差的估算需要使用近似技术。这些技术补充第 2 章与第 3 章中考察过的使用于描述性调查的技术。章节 5.4 将讨论其他基于样本再使用的方法（平衡半样本、折刀以及脱靴），章节 5.5 从数值上来比较这些方法。我们将使用小型芬兰健康调查来演示方差近似法。这一复杂分析性调查使用分层整群抽样，每个地域层级中有两个地域样本整群。章节 5.6 介绍一个更复杂的、职业健康调查（OHC）。OHC 调查的抽样设计是一个一级和两级分层的组合，并以工业实体作为整群。这些数据将用来扩展方差估算到几个比率估计值的协方差矩阵估算。这些比率估计值分别对应具体的子总体。例如，为了建立对数模型和其他类型的建模过程，需要作为子总体比例和均值的比率估计值的协方差矩阵。另一个扩展是讨论非等概选取方法的复杂设计。这是通过整合相应的估计值中的元素权重来实现的。

前面章节介绍的、在复杂抽样设计情形下的比率估计值的方差估算的所有近似方法，将同样适用于协方差矩阵估算。我们选择线性化方法是由于其在实际中的重要性。章节 5.7 将讨论使用线性化的协方差矩阵估算。这里，比率估计值的设计效应的概念被扩展为几个比率估计值向量的设计效应矩阵。在评估整群对协方差矩阵的影响时，也要使用到设计效应矩阵。章节5.8 给出本章小结。

5.1 小型芬兰健康调查

小型芬兰健康调查（MFH 调查——译者注）的设计是为了获取芬兰成年

人健康的总体概览以及他们的照料需求,并发展出在人口中监测健康的方法。这个调查的抽样设计属于两级分层整群抽样的类型。它使用了多种数据收集方法,调查的一个目标是比较这些方法的可靠性(Heliövaara et al. ,1993)。大量的数据是使用移动门诊点的健康检查以及面对面访谈获得。因此,为了费用效率,使用了地域整群的整群抽样。

调查的目标总体是 30 岁及以上的芬兰人。使用的两级分层整群抽样设计中,40 个地理层级中各抽取一个整群。每个层级一个整群的使用,是为了获得在整群总体中更深的分层。以等概选取方法(见章节 3. 2)得到了 8 000 人的样本。回顾一下,等概样本是指使用整体上元素抽样比例相同的设计。

原来的抽样设计

在原来的抽样设计中,320 个人口整群组成了一个自治市,或在某些例子中两个地域上相邻的自治市。这些整群被分层为城镇或农村,以及制造业或是农业的人口份额。从最大的城镇开始,形成 8 个自我呈现的层级。其余 32 个层级由规模相等的整群组成,含有 40 000 ~ 60 000 个符合要求的居民。使用累积方法的 PPS 抽样从这些非确定的层级中抽取一个整群,选中概率与层级中目标总体的规模成比例(见章节 2. 5)。第二级抽样的规模按比例配额得出,因而为等概选取设计。选中整群的样本规模为 50 ~ 500 人,均值为 150 人。个人层次的样本使用登记数据为抽样框,在每个层级中以系统抽样获取,抽样涵盖了整群中相关的人群。

修改后的 MFH 调查抽样设计

由于从每个层级中只抽取一个整群,从不确定层级中无法估算群间方差。因而为了方差估算,使用合并层级技术修改了原来的设计。从 32 不确定层级中生成 16 个虚拟层级,每个新的层级中有两个整群。原有层级中规模大致相等、分层变量取值相似的两个层级合并在一起。为了得到一个便于分析的设计,同时也是为了我们讲解的目的,在 8 个自我呈现的层级中,通过将每个层级中的样本随机分成两个规模近似的部分,生成两个整群。注意,同样的,可以假定这一元素抽样设计在 8 个不确定层级中,每个元素本身就是一个整群,因而修改后的设计有 8 个一级层级和 16 个各自含有 2 个整群的二级层级。成为 MFH 调查抽样设计的修改后的设计,共有 24 个层级和 48 个整群。雷同能与库赛拉(Lehtonen and Kuusela, 1986)对 MFH 设计有更详细的描述。

MFH 调查设计中样本整群相对较小的数目在估算方差与协方差时可能成为问题。整群数目决定方差与协方差的自由度。这些自由度被定义为样本整群数目减去层级数目,即在 MFH 设计中,48 - 24 = 24。较小数目可以造成

方差与协方差的不稳定性,导致检验与建模过程中的问题。在职业健康保健与芬兰健康保险调查中的情形不同,其抽样整群数目较大(章节 5.6 与章节 9.3 将分别描述这些调查)。

数据收集与无应答

实地调查的主要阶段是健康访谈、两个步骤的健康检查,以及一个深度检查。调查在 1978—1981 年实施。主要的方法是访谈、问卷、行为检验、物理和生化测量、观测者评估,以及医生的门诊检查。访谈由当地公共卫生或是医院护士实施,健康检查由移动门诊点实施。

在 8 000 个样本中,7 703 个(96%)完成了健康访谈,7 217 个(90%)参加了健康检查的初选阶段。超过 6 000 个参加初选的个人至少有症状,或是发现或是报告疾病史,因而被要求参加健康检查的门诊阶段;94% 参加了。差不多 5 300 个参加初选检查的个人被要求参加医生的门诊检查;4 840 人参加了。在实地研究后,加入了没参与的数据。因而,所有 5 292 个应邀参加医生检查的个人,均有基于医生检查的门诊数据,或是与这些数据相似的数据。所以,调查的各个阶段应答率非常高。

设计效应

MFH 调查抽样设计中的地域整群含有较大且异质的人群。给定整群的类型,在绝大多数研究变量中只可能有轻微的群内相关系数。但是,也有一些变量有可观的整群效应。表 5.1 给出了一些研究变量样本均值或是比例的设计效应估计值。这些估计值来自健康检查初选阶段的数据。设计效应估计值在 3.2 与 0.9 之间,最大值是连续变量收缩血压。许多研究变量的设计效应估计值接近于 1,而某些估计值小于 1,显示了较弱的整群效应。

表 5.1　MFH 调查数据中一些研究变量样本均值与比例的设计效应估计值

研究变量	deff
收缩血压	3.2
慢性病	2.0
看病次数	1.4
身材胖瘦指数	1.4
血清胆固醇	1.2
看牙医次数	1.0
患病天数	0.9

演示数据

在考查子总体均值和比例的方差近似技术时,我们使用了 MFH 调查数据

的一个部分,30～64 岁参加初选阶段健康检查的男性。他们属于活跃的劳动力或是有过劳动经历。这样的数据由 2 699 个合格男性组成。数据包括抽样标签 STRATUM,CLUSTER 以及 WEIGHT;两个二分因变量,CHRON(有否慢性病)和 PHYS(工作中是否有过身体健康意外);一个连续的因变量 SYSBP(收缩血压)。表 5.2 给出这些数据的信息。注意,选取的样本子群并不局限于某些组别,它恰当地反映了 MFH 调查抽样设计的所有基本特征,诸如涵盖的层级数目(24)和整群数目(48)。

表 5.2　MFH 调查中 30～64 岁男性慢性病患者比例(%),工作有身体健康风险及平均收缩血压的分年龄分布

年　龄	样　本		CHRON	PHYS	SYSBP
	n	%	%	%	Mean
30～34	508	18.8	13.8	12.8	134.0
35～39	384	14.2	21.4	17.4	136.2
40～44	437	16.2	28.4	18.8	138.5
45～49	395	14.6	44.8	18.5	141.9
50～54	379	14.0	52.2	17.4	144.7
55～59	336	12.4	68.5	21.4	151.2
60～64	260	9.6	73.8	21.2	154.3
样本总计	2 699	100.0	39.8	17.8	141.8

我们的目的是,使用基于线性化和样本再使用的近似方法,来估算 CHRON 的子总体比例与 SYSBP 的子总体均值的估计值的方差。从 MFH 调查的整体数据看,两个因变量显示了较强的群内相关。第 7 章在两变量表格的检验中使用因变量 PHYS。在转向这些任务之前,我们简单讨论在相关的 MFH 调查对象中的加权问题。

后续分层

MFH 调查数据可以被看成是自我加权,因为其设计是等概选取,而没必要校正无应答。但是,为了进一步演示章节 3.3 与章节 4.1 中讨论过的后续分层,我们推出后续分层权重,并比较未加权与加权估算结果。为了这一目的,让我们暂时假定当前样本是一个简单随机样本(尽管对于 MFH 调查数据而言,这不是正确的)。

为了得出后续分层权重,我们使用在总体层次已知的两个性别的地域年龄分布。我们首先将目标总体分成 30 个地域年龄—性别后续层级,5 个地域组别,3 个年龄组别。让我们考虑从 MFH 调查选出的 30～64 岁男性;表 5.3 给出相应的总体和样本的频次分布和比例。使用这样的分布,推导出后续层 g 中样本元素的两个权重:权重 $w_g^* = N_g/n_g$ 与换算权重 $w_g^{**} = w_g^* \times n/N$。其

中，N_g 与 n_g 分别是后续层 g 中的总体与样本规模，N 与 n 分别是数据中相应的总体与样本规模。

权重 w_g^* 表示一个样本元素“代表”的总体元素的数目。在 n 个样本的数据中，这些权重加起来等于相应的总体规模 N。换算权重 w_g^{**} 加起来为 n。在表 5.3 中，这些权重仅仅围绕均值 1 轻微波动，显示了 MFH 调查自我加权的特征。在一个实际中并不常见的、严格的自我加权的数据中，权重 w_g^* 是一个常数，而所有样本元素的换算权重 w_g^{**} 均等于 1。

表 5.3 MFH 调查 30～64 岁男性后续分层权重的生成。15 个后续分层的总体与样本规模 N_g 与 n_g，相应的比例 P_g 与 p_g，以及权重 W_g^* 与 w_g^{**}

后续分层	N_g	n_g	P_g	p_g	w_g^*	w_g^{**}
1	56 658	140	0.058 06	0.051 87	404.70	1.119 2
2	32 450	94	0.033 25	0.034 83	345.21	0.954 7
3	21 681	66	0.022 22	0.024 45	328.50	0.908 5
4	71 324	199	0.073 08	0.073 73	358.41	0.991 2
5	41 422	123	0.042 44	0.045 57	336.76	0.931 3
6	33 168	93	0.033 99	0.034 46	356.65	0.986 3
7	75 172	215	0.077 03	0.079 66	349.64	0.966 9
8	45 507	131	0.046 63	0.048 54	347.38	0.960 7
9	33 011	97	0.033 82	0.035 94	340.32	0.941 2
10	116 822	309	0.119 70	0.114 49	378.06	1.045 6
11	62 917	172	0.064 47	0.063 73	365.80	1.011 6
12	47 261	157	0.048 43	0.058 17	301.03	0.832 5
13	188 252	466	0.192 89	0.172 66	403.97	1.117 2
14	88 185	254	0.090 36	0.094 11	347.19	0.960 2
15	62 105	183	0.063 64	0.067 80	339.37	0.938 6
合计	975 935	2 699	1.000 00	1.000 00		

使用权重时，显然 w_g^* 适合于总体总和的估算。而换算权重适合于对总和不感兴趣的检验与建模过程。

在复杂抽样设计中，对于一个非等概选取的数据，生成一个后续分层变量更为复杂。这是因为，可能已有一个补偿非等选中概率的元素权重。在最简单的例子中，可以用后续层应答率乘以抽样权重，得出调整无应答的校正权重，这一乘积可以用在分析程序之中。严格地讲，后续分层估计值的方差估算公式与使用校正权重得到的估算公式不同。但是，在实际中，估计值间的差别很小。

让我们比较从 MFH 调查数据中得出的加权、无加权，以及使用后续分层估算公式的估算结果。为简便起见，我们忽略原有的分层与整群，因而将 MFH 数据在无加权分析中看成是一个简单随机样本（放回式，SRSWR）；在加权分析中，为非比例配额的简单随机分层样本（STRWR）；在第三种情况下，为

简单随机后续分层样本。使用权重 w_g^* 和 w_g^{**} 作为权重变量得到加权估计值；在估算过程中将总体规模 N_g 插入后续层，则实施了后续分层。下面给出相应的 CHRON，PHYS，SYSBP 的样本均值和标准误：

研究变量	n	SRSWR		STRWR		后续分层	
		均值	s. e	均值	s. e	均值	s. e
CHRON	2 699	0. 398	0. 009 4	0. 386	0. 008 4	0. 386	0. 008 5
PHYS	2 699	0. 178	0. 007 4	0. 176	0. 007 3	0. 176	0. 007 3
SYSBP	2 699	141. 8	0. 367 7	141. 4	0. 335 3	141. 4	0. 337 5

CHRON 的无加权与后续分层的均值有差异，SYSBP 的也是如此。这是因为它们对于后续层中人口组成的依赖要强于 PHYS。应当注意到，后续分层可以提高效率。当成简单分层的后续分层分析降低了 CHRON 和 SYSBP 的标准误估计值。由后续分层导致的方差可以从最后一栏中看出（特别对于 SYSBP 而言），这里的标准误估算使用了最合适的方差估算公式。但是，与分层分析相比，差别还是很小。

5.2　比率估算值

在方差估算中，我们集中讨论非线性最简单的例子——比率估计值。总体中次级群体的均值与比例是典型的非线性比率估计值，比如 MFH 调查中收缩血压的均值以及慢性病患者的比例。我们将在分层整群抽样设计的情形下讨论方差估算。这样的设计是与 MFH 调查抽样设计相似的等概选取。对于方差估算，这种抽样设计简单并在实际中广泛使用。

非线性估计值

线性估算公式由一个样本观测值的线性函数组成。从一个预先固定规模的简单随机样本中计算出来的，诸如 $\hat{t} = N\sum_{k=1}^{n} y_k/n$ 的总和是一个线性估计值。在整群抽样中，经常会遇到样本规模无法预先固定的情况。比如，在整群规模 B_i 不同的一级整群抽样中就会发生这样的情况。因而，总和估计值 $\hat{t}_{rat} = N\sum_{i=1}^{m} y_i/\sum_{i=1}^{m} B_i$（章节 3.2 中讨论过）中的分母也应被当成取值取决于抽取的整群的随机变量，其中，y_i 是整群 i 中因变量的样本总和。正因为如此，$\hat{t}_{rat}$ 是一个非线性估计值。

估计值 $\hat{t}_{rat}$ 是章节 3.3 讨论过的比率估算的特例。那里的比率估算是指使

用辅助信息来估算总体总和 T。估计值 $\hat{t}_{rat}=\hat{r}\times T_z$，其中的 $\hat{r}=\hat{t}/\hat{t}_z$ 是因变量与辅助变量 z 的总和估计值 $\hat{t}$ 与 $\hat{t}_z$ 的比率，而 T_z 是 z 的已知总体总和。在总体比率 $R=T/T_z$ 的估算中，估计值 $\hat{r}$ 直接已知，它可以被写成 $\hat{r}=\sum_{i=1}^{m}y_i/\sum_{i=1}^{m}z_i$。估计值 $\hat{r}$ 被称作比率估计值。在上面的估算公式中，分母是并非固定的样本规模。在实际中，从规模不固定的样本中的一个部分估算得出的子总体均值与比例是最常见的比率估算的例子。MFH 调查中的 30～64 岁男性就是这样的部分。我们将讨论这样的比率估算。

组合比率与独立比率估算

将总体整群分成 H 个层级，在层级 h 中得到 M_h 个整群。一级样本从层级 h 中抽取 $m_h(\geqslant 2)$ 个整群，二级样本从 $m=\sum_{h=1}^{H}m_h$ 样本整群中抽取总数为 $n=\sum_{h=1}^{H}n_h$ 个样本元素。由于我们经常使用规模不固定的样本子群，我们将在 n_h 处使用 x_h。注意，我们没有使用 z_h，是为了避免与辅助变量的符号相混淆。我们假定样本是自我加权，亦即 N 个总体元素的选中概率在各个层级中为一常数，没有必要校正无应答。样本中的元素权重相同。让 $y_{hi}=\sum_{k=1}^{x_{hi}}y_{hik}$ 来表示样本层级 h 中的整群 i 的因变量的样本子群的和，x_{hi} 来表示相应的样本规模。使用样本和 y_{hi} 与 x_{hi}，可以推出两种比率估算公式。组合比率估算公式（整体层级比率）为，

$$\hat{r}=\frac{\sum_{h=1}^{H}y_h}{\sum_{h=1}^{H}x_h}=\frac{\sum_{h=1}^{H}\sum_{i=1}^{m_h}y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{m_h}x_{hi}},\tag{5.1}$$

它是均值 $\overline{Y}=T/N$ 或是比例 $P=N_1/N$ 的比率估算公式。其中，T 是连续因变量的总体总和，N_1 是总体中讨论的样本子群在二分因变量取值为 1 的应答者数目。应当注意到，估计值 $\hat{r}$ 的分子 y_{hi} 的数量在整群间不同，而且分母 x_{hi} 的数量也是如此。

在式 5.1 中，y_{hi} 与 x_{hi} 首先在各层级和整群加总。独立比率估算公式（逐个层级比率）是层级比率 y_h/x_h 的加权和，为，

$$\hat{r}_S=\sum_{h=1}^{H}W_h\hat{r}_h,\tag{5.2}$$

其中，$W_h=N_h/N$ 是已知层级权重，且，

$$\hat{r}_h=\frac{y_h}{x_h}=\frac{\sum_{i=1}^{m_h}y_{hi}}{\sum_{i=1}^{m_h}x_{hi}},\ h=1,\cdots,H。$$

独立比率估算公式常用于描述性调查，而组合比率估算公式更常见于复杂的分析性调查中。我们在本章和下一章中，只使用组合比率估算，并将它称为比率估计值。在连续因变量的情形下，我们用 $\hat{r}=\bar{y}$（样本均值）；在二分因变量的情形下，$\hat{r}=\hat{p}$（样本比例）。我们经常将式 5.1 中的比率估计值简写成 $\hat{r}=y/x$，其中，$y=\sum_{h=1}^{H} y_h$，而 $x=\sum_{h=1}^{H} x_h$。y 与 x 的数值分别是样本子群中的因变量的样本和与样本规模。注意，以上的讨论同样适用于从抽样设计不固定规模的整个样本中计算得出的估计值 $\hat{r}$。

变量估计值 $\hat{r}$ 并非无偏，但却是一致的。$\hat{r}$ 的偏差取决于样本中的部分的整群样本规模的变化情况。整群样本规模 x_{hi} 的离异系数可以用来测量这种变化情况。当离异系数较小，比率估计值 $\hat{r}$ 接近于线性，因而接近于无偏。当离异系数小于 0.2 时，偏差并无大碍。

可以生成不同类型的样本子群，其中的比率估计值的偏差特征互不相同。在横断层级和样本整群的整体组群中，整群内部样本子群规模 x_{hi} 的减小与子群规模相对于整个样本规模的减小成比例。子群样本规模的离异系数因而与整个样本的离异系数相同。这样的子群，保持了抽样设计的基本特征，比如，整体组群中的层级和样本整群的数目与整体样本相同。但是，独立组群仅仅涵盖了部分样本整群，子群样本规模的离异系数可能有实质性的增加。地域子群就是这样的例子。应当注意到，与整体组群不同，独立组群并没有准确保留抽样设计的特征，可能导致方差估算中的不稳定性问题（章节 5.7）。在这些极端例子之间是混合组群，它是实际中最常见的子群类型。人口子群通常是整体组群，而社会经济子群则更可能是混合组群。同时，整体组群的子总体比率估计值的设计效应估计值的一个特征是，当子群样本规模降低时，它接近于 1。其他子群类型并没有这一特征。

比率估计值的方差估算

比率估算公式 5.1 中，不仅分子 $\sum_{h=1}^{H} y_h$ 中整群间的离异，而且分母 $\sum_{h=1}^{H} x_h$ 中的离异都生成整体上的方差。因此，比率估计值的方差估算比线性估计值要复杂得多。第 2 章讨论的总体总和的线性估计值的分析性方差估计值，是根据各个基本抽样技术的特征推导出来的。对于非线性估计值而言，分析性方差估计值可能相当复杂，或是根本没有。因而，需要其他类型的方差估计值。为了成功获得这样的估计值，它们及其相应的计算技术应当有着多种目的的特征。这些特征涵盖了最常见的复杂抽样设计与非线性估计值。

可以使用近似方差估计值来估算非线性估计值的方差。与线性估计值不同，这样的估计值并非具体到抽样设计技术。近似方差估计方法非常灵活，可以应用于各种多级设计情形下的各种非线性估计值，包括比率估计值。这些

设计涵盖了本书中所有不同的真实抽样设计。我们使用线性化方法作为基本的近似法。其他方法基于样本再使用技术,包括平衡半样本、折刀,以及脱靴。统计软件中已经含有了复杂调查中方差估算的近似技术。

使用近似方差估计值时,通常设立一些简化的假设。在多级设计的方差估算中,每一个抽样阶段对整个方差各有贡献。比如,在二级设计中,章节3.2显示了,总体总和的分析性方差估计值由群间与群内两个方差部分组成。在近似方法最简单的使用中,潜在的多级设计被简化成一级设计,整群的抽取也被假定使用了放回式的方法。估算方差时仅仅使用群间的离异。在近似技术更高级的使用中,所有抽样阶段的离异均被恰当地纳入其中。

5.3 线性化方法

非线性估计值的线性化方法

在估算由 $\hat{\theta}$ 表示的普通非线性估计值的方差时,我们采用基于所谓泰勒级数展开式的方法。这种方法通常被称作线性化方法。因为,我们首先用相应的泰勒展开式的线性形式,将原来的非线性量简化为近似的线性量,然后改造方差公式和这一线性量的方差估计值。

用 $\mathbf{Y}=(Y_1,\cdots,Y_s)'$来表示一个 s 维的参数向量,其中的 Y_j 是总体总和或均值。相应的估计值向量用 $\hat{\mathbf{Y}}=(\hat{Y}_1,\cdots,\hat{Y}_s)'$来表示,其中,$\hat{Y}_j$ 是 Y_j 的估计值。我们考虑非线性参数 $\theta=f(\mathbf{Y})$,它的一致的估计值表示为 $\hat{\theta}=f(\hat{\mathbf{Y}})$。一个简单的例子是子总体均值参数 $\theta=\overline{Y}=Y_1/Y_2$,其比率估计值为 $\hat{\theta}=\overline{y}=\hat{Y}_1/\hat{Y}_2=y/x$。其中,$y=\sum_{h=1}^{H}\sum_{i=1}^{m_h}y_{hi}$ 是因变量的子群样本和,$x=\sum_{h=1}^{H}\sum_{i=1}^{m_h}x_{hi}$ 是组群样本规模,两者均被当成随机量。

假定对于函数$f(\mathbf{y})$,在含有 $\mathbf{Y}$ 与 $\hat{\mathbf{Y}}$ 的开放空间存在连续的二阶导数。使用泰勒展开式的线性项,我们有近似的线性化表达式,

$$\hat{\theta}-\theta\approx\sum_{j=1}^{s}\frac{\partial f(\mathbf{Y})}{\partial y_j}(\hat{Y}_j-Y_j), \tag{5.3}$$

其中,$\partial f(\mathbf{Y})/\partial y_j$ 是偏微分。使用线性化方程(式5.3),$\hat{\theta}$ 的近似方差可以写成,

$$V(\hat{\theta})\approx V\left(\sum_{j=1}^{s}\frac{\partial f(\mathbf{Y})}{\partial y_j}(\hat{Y}_j-Y_j)\right)=\sum_{j=1}^{s}\sum_{l=1}^{s}\frac{\partial f(\mathbf{Y})}{\partial y_j}\frac{\partial f(\mathbf{Y})}{\partial y_l}V(\hat{Y}_j,\hat{Y}_l), \tag{5.4}$$

其中,$V(\hat{Y}_j,\hat{Y}_l)$表示估计值 $\hat{Y}_j$ 与 $\hat{Y}_l$ 的方差与协方差。因此,我们将非线性估

计值 $\hat{\theta}$ 的方差简化成 s 个线性估计值 $\hat{Y}_j$ 的方差与协方差的函数。使用方差与协方差的估计值 $\hat{v}(\hat{Y}_j, \hat{Y}_l)$ 代替式 5.4 中的 $V(\hat{Y}_j, \hat{Y}_l)$，可以得到方差估计值 $\hat{v}(\hat{\theta})$。得出的方差估计值是一个一阶泰勒展开式的近似值。对高阶项的忽略完全是根据各种样本巨大的复杂调查得出的经验。

作为线性化方法的例子，让我们进一步考察比率估计值。参数向量为 $\mathbf{Y} = (Y_1, Y_2)'$，相应的估计向量为 $\hat{\mathbf{Y}} = (\hat{Y}_1, \hat{Y}_2)'$。将要估计的非线性参数是 $\theta = f(\mathbf{Y}) = Y_1/Y_2$，其相应的比率估计值是 $\hat{\theta} = f(\hat{\mathbf{Y}}) = \hat{Y}_1/\hat{Y}_2$。偏微分是，

$$\partial f(\mathbf{Y})/\partial y_1 = 1/Y_2, \partial f(\mathbf{Y})/\partial y_2 = -Y_1/Y_2^2。$$

因而，我们有，

$$\begin{aligned}
V(\hat{\theta}) &\approx \sum_{j=1}^{2}\sum_{l=1}^{2} \frac{\partial f(\mathbf{Y})}{\partial y_j}\frac{\partial f(\mathbf{Y})}{\partial y_l} V(\hat{Y}_j, \hat{Y}_l) \\
&= \frac{1}{Y_2}\frac{1}{Y_2}V(\hat{Y}_1) + \frac{1}{Y_2}\left(-\frac{Y_1}{Y_2^2}\right)V(\hat{Y}_1, \hat{Y}_2) + \\
&\quad \left(-\frac{Y_1}{Y_2^2}\right)\frac{1}{Y_2}V(\hat{Y}_2, \hat{Y}_1) + \left(-\frac{Y_1}{Y_2^2}\right)\left(-\frac{Y_1}{Y_2^2}\right)V(\hat{Y}_2) \\
&= (1/Y_2^2)[V(\hat{Y}_1) + \theta^2 V(\hat{Y}_2) - 2\theta V(\hat{Y}_1, \hat{Y}_2)] \\
&= \theta^2[Y_1^{-2}V(\hat{Y}_1) + Y_2^{-2}V(\hat{Y}_2) - 2(Y_1Y_2)^{-1}V(\hat{Y}_1, \hat{Y}_2)]。
\end{aligned} \quad (5.5)$$

复杂抽样情形下，非线性估计值的方差估算的线性化方法的基本原则可归功于凯菲茨（Keyfitz, 1957）和泰平（Tepping, 1968）。伍德拉夫（Woodruff, 1971）建议通过将 s 维情形转化为一个单维的情形，来简化近似法的算法。沃尔特（Wolter, 1985）的书是关于这种方法较好的参考书。线性化方法也可用于更复杂的非线性估计值，例如相关系数和回归系数。对于比率估计值和其他更复杂估计值的方差估算，分析软件使用线性化方法。我们接下来讨论使用线性化方法的比率估计值的近似方差的估算。

组合比率估计值的线性化方法

根据等式 5.5，式 5.1 给出的比率估计值 $\hat{r} = y/x = \sum_{h=1}^{H}\sum_{i=1}^{m_h} y_{hi} / \sum_{h=1}^{H}\sum_{i=1}^{m_h} x_{hi}$ 的方差估计值，应当包括以下各项：第一，表示子群样本和 y_{hi} 的整体整群的离异项；第二，表示子群规模 x_{hi} 的整体整群的离异项；最后，表示子群样本和 y_{hi} 与 x_{hi} 的联合整体整群的离异项，即它们的协方差。用估计值 $\hat{v}(y)$，$\hat{v}(x)$ 和 $\hat{v}(y, x)$ 来代替相应的方差和协方差项 $V(y)$，$V(x)$ 和 $V(y,x)$，可以从等式 5.5 中得出 $\hat{r}$ 的方差估计值。因此，我们有，

$$\hat{v}_{des}(\hat{r}) = \hat{r}^2[y^{-2}\hat{v}(y) + x^{-2}\hat{v}(x) - 2(yx)^{-1}\hat{v}(y, x)], \quad (5.6)$$

作为基于线性化方法的、$\hat{r}$ 的基于设计的方差估计值。其中，$\hat{v}(y)$ 是子群样本和 y 的方差估计值，$\hat{v}(x)$ 是子群样本规模 x 的方差估计值，而 $\hat{v}(y,x)$ 是 y 与 x 的协方差。

当估计值 $\hat{v}(y)$，$\hat{v}(x)$ 和 $\hat{v}(y,x)$ 一致时，式 5.6 中的方差估计值也是一致的。由于基于泰勒展开式的近似法的可靠性，整群样本规模 x_{hi} 不应该变动较大。当 x_{hi} 的离异系数小于 0.2 时，可以很安全地使用这个方法。当整群规模相等时，方差与协方差 $\hat{v}(x)$ 和 $\hat{v}(y,x)$ 为 0，而方差近似值简化为 $\hat{v}_{des}(\hat{r}) = \hat{v}(y)/x^2$。对于放回式简单随机抽样得来的二分变量，这一方差估计值简化为二项方差估计值 $\hat{v}_{des}(\hat{p}) = \hat{v}_{bin}(\hat{p}) = \hat{p}(1-\hat{p})/x$。其中，$x=n$，它是已有的样本数据的规模。

当元素层次的样本规模较大，并且样本整群数目较大时，方差估计值公式 5.6 可以得到较好的方差估计值。因此，它是一个大样本近似。在样本整群数目较小时，方差估计值可能并不稳定。章节 5.7 将考察这样的情形。

严格地讲，式 5.6 中的方差和协方差估计值取决于实际的抽样设计。但是，假定每个层级中至少抽取两个整群，并且使用放回式的假设，亦即假定整群的抽取是相互独立的，那么，我们得到相对简单的方差和协方差估计值。它们也可用于多级分层的等概抽取样本：

$$\hat{v}(y) = \sum_{h=1}^{H} m_h \hat{s}_{yh}^2, \quad \hat{v}(x) = \sum_{h=1}^{H} m_h \hat{s}_{xh}^2$$

及

$$\hat{v}(y,x) = \sum_{h=1}^{H} m_h \hat{s}_{yxh},$$

其中，

$$\hat{s}_{yh}^2 = \sum_{i=1}^{m_h} (y_{hi} - y_h/m_h)^2/(m_h - 1),$$

$$\hat{s}_{xh}^2 = \sum_{i=1}^{m_h} (x_{hi} - x_h/m_h)^2/(m_h - 1),$$

以及

$$\hat{s}_{yxh} = \sum_{i=1}^{m_h} (y_{hi} - y_h/m_h)(x_{hi} - x_h/m_h)/(m_h - 1)。 \quad (5.7)$$

注意，由于使用放回式近似，只考虑了群间的离异。因此，相应的方差估计值低估了真实的方差。当整体层级的第一级抽样比例较小时，这一偏差可以忽略。这其实就是每个层级中总体整群数目较大的情形（见章节 3.2）。

为了估算群间的方差，至少需要两个样本整群。当样本设计在各个层级中仅抽取两个整群时，式 5.7 中的估算公式可以进一步简化为，

$$\hat{v}(y) = \sum_{h=1}^{H} (y_{h1} - y_{h2})^2, \quad \hat{v}(x) = \sum_{h=1}^{H} (x_{h1} - x_{h2})^2$$

以及

$$\hat{v}(y,x) = \sum_{h=1}^{H} (y_{h1} - y_{h2})(x_{h1} - x_{h2})。 \quad (5.8)$$

由于其简单的方差和协方差估算公式，这种设计在实际很常见。修改后的MFH 调查抽样设计就是这种类型。范例 5.1 演示了 MFH 调查二级设计的线性化方法。

范例 5.1

MFH 调查的线性化方法。我们使用线性化方法来讨论二分因变量 CHRON（慢性病）子总体比例估计值 $\hat{r}=\hat{p}$ 以及连续因变量 SYSBP（收缩血压）子总体均值估计值 $\hat{r}=\bar{y}$ 的方差估算。MFH 调查的子群包括 30～64 岁男性。子群的样本规模为 $x=2\ 699$，数据是自我加权的。章节 5.1 描述过的修改的 MFH 调查抽样设计有 $H=24$ 个地域层级和 $m=48$ 个地域样本整群。每个层级中抽取了两个整群。回忆一下，由于子群组成了一个整体类型的组群，因而它保持了这些抽样设计的特征。表 5.4 给出了数据。

表 5.4 MFH 调查中回应变量 CHRON 与 SYSBP 的整群样本和 y_{hi} 与相应的 30～64 岁男性整群样本规模 x_{hi}

层级	整群	CHRON	SYSBP		整群	CHRON	SYSBP	
h	i	y_{hi}	y_{hi}	x_{hi}	i	y_{hi}	y_{hi}	x_{hi}
1	1	70	29 056	204	2	74	29 417	210
2	1	12	3 692	26	2	14	4 564	30
3	1	15	7 741	59	2	16	8 585	63
4	1	9	6 277	45	2	14	5 668	43
5	1	10	2 322	17	2	16	3 960	30
6	1	10	3 080	21	2	6	3 252	22
7	1	10	3 966	27	2	4	3 261	24
8	1	12	4 156	28	2	6	2 852	20
9	1	15	6 617	46	2	23	6 616	48
10	1	37	10 552	73	2	25	11 032	77
11	1	11	8 759	60	2	25	9 876	72
12	1	33	9 901	69	2	24	6 828	47
13	1	31	8 624	61	2	27	9 390	66
14	1	22	6 960	48	2	20	7 130	49
15	1	18	6 646	49	2	22	7 094	49
16	1	24	9 841	69	2	37	11 786	83
17	1	19	6 910	48	2	23	6 446	45
18	1	25	10 742	73	2	29	9 026	61
19	1	36	9 350	65	2	34	8 912	62
20	1	9	3 810	26	2	22	7 098	51
21	1	18	6 998	53	2	34	9 970	69
22	1	29	11 146	79	2	41	13 215	94
23	1	22	6 596	48	2	18	6 002	41
24	1	15	3 808	27	2	7	3 148	22
两个整群所有层级加总						1 073	382 678	2 699

对于二分因变量 CHRON,我们有

$$y = \sum_{h=1}^{24} \sum_{i=1}^{2} y_{hi} = \sum_{h=1}^{24} (y_{h1} + y_{h2}) = 1\ 073$$

个患有慢性病的男性,子群中的样本和为,

$$x = \sum_{h=1}^{24} \sum_{i=1}^{2} x_{hi} = \sum_{h=1}^{24} (x_{h1} + x_{h2}) = 2\ 699$$

个男性。CHRON 的子总体比例估计值为,

$$\hat{p} = y/x = 1\ 073/2\ 699 = 0.397\ 6。$$

对于 $\hat{p}$ 的方差估计值 $\hat{v}_{des}(\hat{p})$,我们计算方差和协方差 $\hat{v}(y)$,$\hat{v}(x)$ 和 $\hat{v}(y,x)$。使用等式 5.8,它们是,

$$\hat{v}(y) = \sum_{h=1}^{24} (y_{h1} - y_{h2})^2 = 1\ 545,\ \hat{v}(x) = \sum_{h=1}^{24} (x_{h1} - x_{h2})^2 = 2\ 527$$

以及

$$\hat{v}(y,x) = \sum_{h=1}^{24} (y_{h1} - y_{h2})(x_{h1} - x_{h2}) = 1\ 435。$$

使用这些估计值,我们得到式 5.6 的方差估计值,

$$\begin{aligned}\hat{v}_{des}(\hat{p}) &= \hat{p}^2[y^{-2}\hat{v}(y) + x^{-2}\hat{v}(x) - 2(y \times x)^{-1}\hat{v}(y,x)] \\ &= 0.397\ 6^2 \times [1\ 073^{-2} \times 1\ 545 + 2\ 699^{-2} \times 2\ 527 - \\ &\quad 2 \times (1\ 073 \times 2\ 699)^{-1} \times 1\ 435] = 0.110\ 3 \times 10^{-3}。\end{aligned}$$

对于连续因变量 SYSBP,我们得到样本和

$$y = \sum_{h=1}^{24} \sum_{i=1}^{2} y_{hi} = \sum_{h=1}^{24} (y_{h1} + y_{h2}) = 382\ 678。$$

因此,SYSBP 的子总体均值估计值是,

$$\bar{y} = y/x = 382\ 678/2\ 699 = 141.785。$$

对于 $\bar{y}$ 的方差估计值 $\hat{v}_{des}(\bar{y})$,我们得到,

$$\hat{v}(y) = \sum_{h=1}^{24} (y_{h1} - y_{h2})^2 = 50\ 469\ 516$$

及

$$\hat{v}(y,x) = \sum_{h=1}^{24} (y_{h1} - y_{h2})(x_{h1} - x_{h2}) = 349\ 962。$$

使用这些估计值,我们得到式 5.6 的方差估计值,

$$\begin{aligned}\hat{v}_{des}(\bar{y}) &= \bar{y}^2[y^{-2}\bar{v}(y) + x^{-2}\hat{v}(x) - 2(y \times x)^{-1}\hat{v}(y,x)] \\ &= 141.785^2 \times [382\ 678^{-2} \times 50\ 469\ 516 + 2\ 699^{-2} \times 2\ 527 - \\ &\quad 2 \times (382\ 678 \times 2\ 699)^{-1} \times 349\ 962] = 0.278\ 8。\end{aligned}$$

所有这些方差可以从表 5.4 给出的整群层次的数据中估算出来。对于 CHRON,我们接下来计算对应于放回式简单随机抽样的 $\hat{p}$ 的二项方差估计值,以及相应的设计效应估计值。方差估计值为,

$$\hat{v}_{bin}(\hat{p}) = \frac{\hat{p}(1 - \hat{p})}{x}$$

$$=0.3976\times(1-0.3976)/2\,699=0.0887\times10^{-3},$$

其中,$\hat{v}_{bin}$是标准的二项方差估计值。设计效应估计值是 $\hat{d}(\hat{p})=\hat{v}_{des}(\hat{p})/\hat{v}_{bin}(\hat{p})=1.24$。注意,由于子群是横跨层级的,这一设计效应估计值比整个调查数据的设计效应小很多。这个估计值同时显示,子群中 CHRON 的群内相关系数很小。另一方面,对于 SYSBP,计算放回式简单随机抽样情形下的 $\bar{y}$ 的方差估计值,需要个人层次的数据。方差估计值为,

$$\hat{v}_{srswr}(\bar{y})=\sum_{k=1}^{2\,699}(y_k-\bar{y})^2/[2\,699(2\,699-1)]=0.1352,$$

而设计效应估计值为 $\hat{d}(\bar{y})=2.06$。它显示了,即使它比整个调查数据的设计效应估计值小很多,组群中因变量 SYSBP 的群内相关系数却不小。子群样本规模的离异系数为,$c.v(x)=s.e(x)/x=0.019$,小到足够合理使用泰勒展开式线性化。最后,我们将估算结果整理如下。

研究变量	估计值	标准误估计值		deff
		$s.e_{des}(\hat{r})$	$s.e_{srs}(\hat{r})$	
CHRON	0.397 6	0.010 5	0.009 4	1.24
SYSBP	141.785	0.528 0	0.367 7	2.06

在实际中,比率类型的比例或是均值估计值的方差估算可以用合适的调查分析软件来处理。调查分析软件使用的是个人层次的数据,而非整群层次的数据。需要进一步训练的读者,可以参考本书的扩展网页。

5.4　样本再使用方法

在复杂多级设计的情形下,样本再使用方法可以当成是非线性估计值 $\hat{\theta}$ 的方差近似估算的线性化方法的替代。术语再使用是指,方差估算是根据样本数据的反复使用。而样本本身仅仅是从总体中获得的单一样本。所以,这些方法有时被称为虚拟重复技术。应当将虚拟重复与诸如随机群方法区分开来。后者依赖于真正的重复——从同一个总体里抽取几个独立的样本。由于它们在复杂分析性调查中有限的实用性,这里并不包括这样的方法。

在这一小节中,我们将讨论 3 种特别的样本再使用技术:平衡半样本、折刀及脱靴。它们有着以下共同的方差估算程序(它们事实上起源于随机群方法):

1. 从样本数据中,我们使用含有数值 K 的具体的技术抽取 K 个虚拟样

本。各个再使用方法各不相同。

2. 从这 K 个样本中获取模仿原估计值 $\hat{\theta}$ 的估计值 $\hat{\theta}_k$。

3. 使用观测到的虚拟样本估计值 $\hat{\theta}_k$ 的离异，特别是根据 $(\hat{\theta}_k - \hat{\theta})^2$ 的平方差来估算估计值 $\hat{\theta}$ 的方差 $V(\hat{\theta})$。典型的，样本再使用估计值的形式为 $\hat{v}(\hat{\theta}) = c\sum_{k=1}^{K}(\hat{\theta}_k - \hat{\theta})^2$。其中，$c$ 为常数，各种再使用方法各不相同。在求平方差时，$\dot{K}$ 个虚拟样本估计值 $\hat{\theta}_k$ 的平均值 $\bar{\hat{\theta}} = \sum_{k=1}^{K}\hat{\theta}_k/K$ 可以用来代替 $\hat{\theta}$。

估计值 $\hat{\theta}$ 通常是一个非线性估计值—— 一个比率估计值或是回归系数估计值。在线性化方法中，构建方差估算公式时，需要这样的非线性方程偏微分的分析性表达式。样本再使用技术却并不一定如此。事实上，上面描述的基本的方差估算程序与估计值的类型无关。因此，方法是适用于任意类型的非线性估计值的。但是，虚拟重复技术，特别是脱靴，比线性化方法牵涉到更多的计算。因此，它们虽然灵活，但是运算繁复。

平衡半样本由麦卡锡(McCarthy, 1966, 1969)首先介绍，用来近似估算在等概抽取设计下的非线性估计值的方差。这样的设计含有大量的层级，而在每个层级中用放回式方法恰好抽取两个整群。在相似的设计中，麦卡锡(McCarthy, 1966)还介绍了折刀方法。此一方法最早由昆诺尔(Quenouille, 1956)在讨论降低估计值偏差时发展出来。折刀方法的一个关键特征的精要概述是行行皆通，样样稀松。两种方法都扩展到了更复杂的设计，包括各层级可以有超过两个的整群及无放回式抽取整群。沃尔特(Wolter, 1985)和拉奥等(Rao et al., 1992)是介绍复杂调查里平衡半样本和折刀技术的很好的文献。

脱靴由埃弗龙(Efron, 1982)以解决各种统计问题的通用的非参数方法引入的，"我们的目标是理解一系列关于偏差、方差及更广泛的误差指标的非参数估算的思想"(Efron,1982,第 1 页)。从此，通过使用繁复的计算机模拟，这一技术广泛地用于独立观测值的各种非标准方差和置信区间的近似估算中。与折刀方法一样，脱靴技术也生成于调查抽样的框架之外，它仅仅是在近年使用于复杂调查非线性估计值的方差估算之中。麦卡锡与斯诺登(McCarthy and Snowden, 1985)是首先将它用于无放回式固定总体抽样的先驱之一。拉奥与吴(Rao and Wu, 1988)，拉奥等(Rao et al., 1992)以及西特尔(Sitter, 1992,1997)广泛地讨论了脱靴技术，包括诸如分位值的非光滑函数。桑德尔等(Särndal et al., 1992)给出了复杂调查中脱靴技术的简要小结。

在这里，我们只介绍样本再使用的基本原则，并专注于它们在 MFH 调查背景下的实际使用。我们将再次讨论式 5.1 给出的作为非线性估计值例子的

(组合)比率估计值$\hat{r}=y/x$。其中,$y=\sum_{h=1}^{H}\sum_{i=1}^{m_h}y_{hi}$是因变量子群样本和在整群层次的和,$x=\sum_{h=1}^{H}\sum_{i=1}^{m_h}x_{hi}$是相应的子群样本规模在整群层次的和。假定放回式抽取整群的二级等概选取抽样设计。放回式的假设将在近似方差估计值中生成偏差,但是,当一级抽样比例较小时,这一偏差可以忽略。注意,再使用方法方差近似估算中的整群层次的数据与线性化方法中使用的相似。

我们将在每个层级恰好抽取两个整群的设计中来检视比率估计值$\hat{r}$近似估算的平衡半样本与折刀技术。注意,MFH 调查抽样设计正是这一类型。使用脱靴技术的是一个在各个层级中至少抽取两个整群,但样本整群数目为常数的更普遍的设计。在这些设计中,这些技术被称作平衡重复复制(BRR)、折刀重复复制(JRR)、脱靴重复复制(BOOT)。因为在文献中,有其他几种 BRR 与 JRR 的版本,我们的目的也是比较各种估算结果,同时也与线性化方法的结果相比较。章节 5.5 给出整体比较。

样本再使用在它们的渐进性和其他特征上——计算要求与实用性——各不相同。基什与弗兰克尔(Kish and Frankel, 1970,1974)、比恩(Bean, 1975)、克鲁斯基与拉奥(Krewski and Rao, 1981)、拉奥与吴(Rao and Wu, 1985, 1988)、拉奥等(Rao et al., 1992)、邵与涂(Shao and Tu, 1995)都报告了复杂调查中使用样本再使用方法估算非线性估计值的结果比较。我们将在章节 5.5 中简要讨论这些方法的相对优势。

BRR 技术

就其基本形式而言,平衡重复复制技术可以用于每个层级恰好用放回式方法抽取两个整群而层级数目较大的等概抽取设计的方差近似估算。通过这种设计,我们将讨论在作为子总体均值或是比例估计值的比率估计值$\hat{r}=y/x$中使用 BRR 方法。其中,$y=\sum_{h=1}^{H}(y_{h1}+y_{h2})$,$x=\sum_{h=1}^{H}(x_{h1}+x_{h2})$,而$y_{hi}$与$x_{hi}$($i=1,2$)是前面给出的整群层次的样本和。

BRR 技术中构造虚拟样本的起点是,给定H个层级与每个层级$m_h=2$样本整群,整体样本可以分成2^H个相互交叉的半样本,每一个这样的样本含有H个样本整群。对于每一个半样本,从第一个层级中选取(y_{11},x_{11})与(y_{12},x_{12}),从第二个层级中选取(y_{21},x_{21})与(y_{22},x_{22}),如此等等。每个半样本k,有比率估计公式,

$$\hat{r}_k=\frac{\sum_{h=1}^{H}\sum_{i=1}^{2}\delta_{hik}y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{2}\delta_{hik}x_{hi}},\ k=1,\cdots,2^H \tag{5.9}$$

其中,整群hi从第k个半样本选取时,权重$\delta_{hik}=1$,其余$\delta_{hik}=0$。

所有半样本中，均值 $\hat{r}_k$ 为，

$$\bar{\hat{r}} = \sum_{k=1}^{2^H} \hat{r}_k / 2^H, \tag{5.10}$$

它和源估计值 $\hat{r}$ 的方差估计值可以用半样本中得出的 $\hat{r}_k$ 构造出来。我们有，

$$\hat{v}(\bar{\hat{r}}) = \sum_{k=1}^{2^H} (\hat{r}_k - \bar{\hat{r}})^2 / 2^H$$

及

$$\hat{v}(\hat{r}) = \sum_{k=1}^{2^H} (\hat{r}_k - \hat{r})^2 / 2^H。 \tag{5.11}$$

如果 $\hat{r}$ 是线性估计值，恒等式 $\hat{r} = \bar{\hat{r}}$ 成立，并且式 5.11 中的两个方差估计值相等。尽管对于比率估计值，这一等式不成立，但是，实际中源估计值与半样本估计值的均值通常非常接近。所以，式 5.11 中任何一个方差估计值均可以用作源估计值 $\hat{r}$ 的方差估计值。显然，这样的方差估算公式并不实用，因为它们假定了数目巨大的半样本，例如 MFH 调查背景下的 1 700 万。为了避免构造所有虚拟样本的繁复任务，可以从中选取一个子集。但是，如果这一子集是随机抽取，则相应的方差估算公式中，会出现一项层级间的协方差。在 BRR 技术中，用平衡方法选取一个 K 个半样本的子集。平衡是指选取半样本时使得层级间的协方差为 0。这样就减少了所需的半样本数量。实际中，数目 K 至少应与层级数目 H 相等。

平衡选取半样本的使用方法，由普莱克特与伯曼（Plackett and Burman, 1946）在构造 K 为 4 的正整倍数的 $K \times K$ 正交矩阵时发展出来。下面将给出这样的正交哈达玛德矩阵 $\mathbf{B}$ 的例子，$K = 12$，$\mathbf{B}'\mathbf{B} = 12 \times \mathbf{I}$，$\mathbf{I}$ 表示单位矩阵。矩阵的行是半样本，列是层级。矩阵中元素 (k,h) 中的 +1 表示层级 h 中的第一个整群 $h1$ 包含在第 k 个半样本中，而 -1 则表示第二个整群 $h2$ 包含在其中。注意，补充半样本可以通过简单地改变矩阵中的符号而得到。半样本的数目，$K = 12$，大大小于在这里为 $2^{12} = 4\ 096$ 的潜在半样本的总数。

半样本 K	层级 h											
	1	2	3	4	5	6	7	8	9	10	11	12
1	+1	-1	+1	-1	-1	-1	+1	+1	+1	-1	+1	-1
2	+1	+1	-1	+1	-1	-1	-1	+1	+1	+1	-1	-1
3	-1	+1	+1	-1	+1	-1	-1	-1	+1	+1	+1	-1
4	+1	-1	+1	+1	-1	+1	-1	-1	-1	+1	+1	-1
5	+1	+1	-1	+1	+1	-1	+1	-1	-1	-1	+1	-1
6	+1	+1	+1	-1	+1	+1	-1	+1	-1	-1	-1	-1
7	-1	+1	+1	+1	-1	+1	+1	-1	+1	-1	-1	-1
8	-1	-1	+1	+1	+1	-1	+1	+1	-1	+1	-1	-1
9	-1	-1	-1	+1	+1	+1	-1	+1	+1	-1	+1	-1
10	+1	+1	-1	-1	+1	+1	+1	-1	+1	+1	-1	-1
11	-1	-1	-1	-1	+1	+1	+1	-1	+1	+1	+1	-1
12	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

如果层级的实际数目为12,我们在平衡构造半样本时使用整个方阵。若 H 小于 K,如10,我们则使用矩阵的任意10行。在 MFH 调查设计中,我们将使用 $K=24$,它等于层级数目。估算线性估计值时,通过选择 K 是大于 H 的4的整数倍数,达成了全正交平衡。这涉及以半样本估计值的均值为整个样本均值估计值的等式。沃尔特(Wolter,1985)给出了秩为2~100的哈达玛德矩阵;使用合适的计算机运算法则,也可以简易地得到这样的矩阵。

文献中给出了几个源估计值 $\hat{r}$ 的方差 $V(\hat{r})$ 的 BRR 估计值。基于 K 个半样本估计值 $\hat{r}_k$ 与全样本估计值 $\hat{r}$ 的方差估算公式为,

$$\hat{v}_{1.brr}(\hat{r}) = \sum_{k=1}^{K}(\hat{r}_k - \hat{r})^2/K, \tag{5.12}$$

它与基于整个 2^H 的式5.11相等。对应于方差估计值 $\hat{v}_{1.brr}(\hat{r})$ 的、基于从 K 个补充半样本中得出的估计值 $\hat{r}_k^C$ 的估算公式为,

$$\hat{v}_{2.brr}(\hat{r}) = \sum_{k=1}^{K}(\hat{r}_k^C - \hat{r})^2/K。 \tag{5.13}$$

式5.12和式5.13使用方差估计值得出组合方差估计值,

$$\hat{v}_{3.brr}(\hat{r}) = (\hat{v}_{1.brr}(\hat{r}) + \hat{v}_{2.brr}(\hat{r}))/2 \tag{5.14}$$

根据 $\hat{r}_k$ 与 $\hat{r}_k^C$ 的均值可以得出的式5.12,式5.14的对应方差估计值,与 $\hat{v}_{1.brr}$ 相应的估算公式为,

$$\hat{v}_{4.brr}(\hat{r}) = \sum_{k=1}^{K}(\hat{r}_k - \bar{\hat{r}})^2/K, \text{其中} \bar{\hat{r}} = \sum_{k=1}^{K}\hat{r}_k/K, \tag{5.15}$$

使用补充半样本构造的估算公式为,

$$\hat{r}_{5.brr}(\hat{r}) = \sum_{k=1}^{K}(\hat{r}_k^C - \bar{\hat{r}}^C)^2/K, \text{其中} \bar{\hat{r}}^C = \sum_{k=1}^{K}\hat{r}_k^C/K。 \tag{5.16}$$

使用 $\hat{v}_{4.brr}$ 与 $\hat{v}_{5.brr}$,我们得到与 $\hat{v}_{3.brr}$ 相对的,

$$\hat{v}_{6.brr}(\hat{r}) = (\hat{v}_{4.brr}(\hat{r}) + \hat{v}_{5.brr}(\hat{r}))/2。 \tag{5.17}$$

使用所有半样本中的 $\hat{r}_k$ 与 $\hat{r}_k^C$,我们最后有,

$$\hat{v}_{7.brr}(\hat{r}) = \sum_{k=1}^{K}(\hat{r}_k - \hat{r}_k^C)^2/4K。 \tag{5.18}$$

对于线性估计值,所有这些方差估算公式相吻合。但是,对于比率估计值,并非如此。比如,$\hat{v}_{3.brr}$ 与 $\hat{v}_{7.brr}$ 之间有着关系,

$$\hat{v}_{3.brr}(\hat{r}) = \hat{v}_{7.brr}(\hat{r}) + \sum_{k=1}^{K}(\bar{\hat{r}} - \hat{r})^2/K,$$

因而,$\hat{v}_{3.brr}(\hat{r}) \geqslant \hat{v}_{7.brr}(\hat{r})$。根据沃尔特(Wolter, 1985),$\hat{v}_{7.brr}$ 可以看成是源估计值 $\hat{\theta}$ 最自然的 BRR 方差估计值。但在实际中,正如 MFH 调查表现出来的,所有估算公式得出近乎相同的方差估计值。

范例 5.2

MFH 调查中的 BRR 技术。与前面小节讨论的线性化方法相同,我们继

续讨论 MFH 调查数据中比率类型子总体均值和比例估计值的方差的近似估算。我们使用二项因变量 CHRON(慢性病)和连续因变量 SYSBP(收缩血压)。子集由 30～64 岁男性组成,其规模为 2 699。CHRON 的比例估计值用 $\hat{r}=\hat{p}$ 来表示,SYSBP 的均值估计值用 $\hat{r}=\bar{y}$ 来表示。我们计算所有 7 个 $\hat{p}$ 与 $\bar{y}$ 的 BRR 方差估计值。

回忆一下,修改后的 MFH 设计有 $H=24$ 个层级与 $m=48$ 个样本整群,恰好每个层级两个整群。BRR 估算开始于构造 K 个半样本和相应的补充半样本。我们选择 $K=24$,即层级数目,并使用整个矩阵来构造半样本以及它们的补充半样本。注意,为了全正交平衡,我们可以选择 $K=28$。我们从 24×24 哈达玛德矩阵中得出用于计算的权重矩阵,它是根据范例 5.1 中给出的整群层次的数据得出的。

首先计算源比率和均值估计值 $\hat{p}$ 与 $\bar{y}$,以及相应的、含有它们各自补充半样本估计值 $\hat{p}_k^C$ 与 $\bar{y}_k^C$ 的半样本估计值均值 $\hat{p}_k$ 与 $\bar{y}_k$。它们是,

$$\hat{p}=0.397\,6,\ \bar{\hat{p}}=\sum_{k=1}^{24}\hat{p}_k/24=0.395\,3\quad 与\quad \bar{\hat{p}}^C=\sum_{k=1}^{24}\hat{p}_k^C/24=0.399\,7,$$

$$\bar{y}=141.785,\ \hat{\bar{y}}=\sum_{k=1}^{24}\bar{y}_k/24=141.804\quad 与\quad \hat{\bar{y}}^C=\sum_{k=1}^{24}\bar{y}_k^C/24=141.768。$$

CHRON 的 3 个比例估计值以及 SYSBP 的 3 个均值估计值都十分接近。我们接下来计算 BRR 方差估计值(式 5.12 至式 5.18)。对于 CHRON,使用 $\hat{p}$,我们从半样本及它们的补充样本中得到,

$$\hat{v}_{1.brr}(\hat{p})=\sum_{k=1}^{24}(\hat{p}_k-0.397\,6)^2/24=0.110\,4\times10^{-3},$$

$$\hat{v}_{2.brr}(\hat{p})=\sum_{k=1}^{24}(\hat{p}_k^C-0.397\,6)^2/24=0.110\,3\times10^{-3},$$

及

$$\hat{v}_{3.brr}(\hat{p})=(\hat{v}_{1.brr}(\hat{p})+\hat{v}_{2.brr}(\hat{p}))/2=0.110\,3\times10^{-3}。$$

使用均值估计值$\bar{\hat{p}}$与$\bar{\hat{p}}^C$,我们得到对应估计值,

$$\hat{v}_{4.brr}(\hat{p})=\sum_{k=1}^{24}(\hat{p}_k-0.395\,3)^2/24=0.105\,2\times10^{-3},$$

$$\hat{v}_{5.brr}(\hat{p})=\sum_{k=1}^{24}(\hat{p}_k^C-0.399\,7)^2/24=0.105\,6\times10^{-3},$$

及

$$\hat{v}_{6.brr}(\hat{p})=(\hat{v}_{4.brr}(\hat{p})+\hat{v}_{5.brr}(\hat{p}))/2=0.105\,4\times10^{-3}。$$

从所有半样本中,我们最后得到,

$$\hat{v}_{7.brr}(\hat{p})=\sum_{k=1}^{24}(\hat{p}_k-\hat{p}_k^C)^2/(4\times24)=0.110\,3\times10^{-3}。$$

对于 CHRON 而言,前面 3 个及最后一个 BRR 估计值与线性化方法得到的方差估计值相等。根据半样本估计值均值的方差估计值稍微——但不是太过地——小一点。

对于 SYSBP,我们得到以下 BRR 方差估计值,

$$\hat{v}_{1.brr}(\bar{y}) = \sum_{k=1}^{24}(\bar{y}_k - 141.785)^2/24 = 0.2791,$$

$$\hat{v}_{2.brr}(\bar{y}) = \sum_{k=1}^{24}(\bar{y}_k^C - 141.785)^2/24 = 0.2790,$$

$$\hat{v}_{3.brr}(\bar{y}) = (\hat{v}_{1.brr}(\bar{y}) + \hat{v}_{2.brr}(\bar{y}))/2 = 0.2791,$$

$$\hat{v}_{4.brr}(\bar{y}) = \sum_{k=1}^{24}(\bar{y}_k - 141.804)^2/24 = 0.2787,$$

$$\hat{v}_{5.brr}(\bar{y}) = \sum_{k=1}^{24}(\bar{y}_k^C - 141.768)^2/24 = 0.2788,$$

$$\hat{v}_{6.brr}(\bar{y}) = (\hat{v}_{4.brr}(\bar{y}) + \hat{v}_{5.brr}(\bar{y}))/2 = 0.2787,$$

$$\hat{v}_{7.brr}(\bar{y}) = \sum_{k=1}^{24}(\bar{y}_k - \bar{y}_k^C)^2/(4 \times 24) = 0.2790。$$

当四舍五入到 3 位小数点时,所有 SYSBP 的 BRR(以及线性化方法得到的)方差估计值均为 0.279。

对于讨论过的因变量而言,所有 BRR 方差估计值得出了相近的子总体比例或均值的比率估计值的结果。这样的结果与其他相似检验研究的结果相同。即使在理论上,无法对非线性估计值的 BRR 方差估算公式给出确定的偏好。除了这里介绍的 BRR 估算公式,也有其他版本。比如称作费伊方法(Judkins,1990)的 BRR 变种,与折刀类型的估算非常相似。

JRR 技术

由于仅仅在构造虚拟样本的方法上有所不同,基于折刀重复复制的折刀方法含有很多 BRR 技术的特点。与 BRR 相比,JRR 技术应用于每个层级抽取两个以上整群的设计的过程更加直截了当。但是,我们仅仅讨论最简单的情形,每个层级恰好抽取两个整群,并假定放回式的形式,即,与 BRR 所要求的设计相类似。我们将推导表现为子总体比例或均值估计值的比率估计值的 JRR 方差估算公式。

我们使用弗兰克尔(Frankel, 1971)建议的方法来构造虚拟样本。对于第一个虚拟样本,我们排除第一层级中的第一个整群 $h1$,并用 2 加权第二个整群 $h2$,而余下的 $H-1$ 个层级保留不变。重复这样的程序,我们共得到 H 个虚拟样本。对于相似的 H 个补充虚拟样本,我们转换排除整群的顺序。JRR 方差估算公式由这两组虚拟样本推导而出。

与 BRR 技术一样,源比率估计值 $\hat{r}$ 的其余几个 JRR 方差估算公式可以构造出来。为了达成这一目的,我们首先推出各个层级的虚拟样本估计值。用 $\hat{r}_h$ 来表示层级 h 中排除整群 $h1$ 并复制整群 $h2$ 的虚拟样本估计值,

$$\hat{r}_h = \frac{2y_{h2} + \sum_{h' \neq h}^{H} \sum_{i=1}^{2} y_{h'i}}{2x_{h2} + \sum_{h' \neq h}^{H} \sum_{i=1}^{2} x_{h'i}}, \ h = 1, \cdots, H。 \tag{5.19}$$

每个虚拟样本都有各自的这样的估计值。从补充虚拟样本中，通过排除整群 $h2$、复制整群 $h1$，我们得到相应的估计值 $\hat{r}_h^C$。使用虚拟样本估计值以及补充虚拟样本估计值，我们可以推出源估计值 $\hat{r}$ 的第一组 JRR 方差估算公式。所以，我们有，

$$\hat{v}_{1.jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h - \hat{r})^2, \tag{5.20}$$

从补充虚拟样本中，还有，

$$\hat{v}_{2.jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h^C - \hat{r})^2。 \tag{5.21}$$

组合方差估算公式为，

$$\hat{v}_{3.jrr}(\hat{r}) = (\hat{v}_{1.jrr}(\hat{r}) + \hat{v}_{2.jrr}(\hat{r}))/2。 \tag{5.22}$$

另一组方差估算公式可以使用所谓的虚拟值来获得。这一概念由昆诺尔(Quenouille, 1956)引入，用来减少估计值偏差。在上述讨论的情形里，虚拟值的形式为，

$$\hat{r}_h^p = 2\hat{r} - \hat{r}_h, \ h = 1, \cdots, H, \tag{5.23}$$

对于补充虚拟样本而言，其符号为 $\hat{r}_h^{pc}$。使用第一组 H 个虚拟值 $\hat{r}_h^p$，我们得到偏差校正估计值，

$$\bar{\hat{r}}^p = \sum_{h=1}^{H} \hat{r}_h^p / H, \tag{5.24}$$

使用补充虚拟样本中的虚拟值 $\hat{r}_h^{pc}$，我们得到，

$$\bar{\hat{r}}^{pc} = \sum_{h=1}^{H} \hat{r}_h^{pc} / H。 \tag{5.25}$$

方差估算公式 5.20 到式 5.22 的对应公式，可以从虚拟值和偏差校正估计值得出，

$$\hat{v}_{4.jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h^p - \bar{\hat{r}}^p)^2, \tag{5.26}$$

从补充虚拟样本，有，

$$\hat{v}_{5.jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h^{pc} - \bar{\hat{r}}^{pc})^2。 \tag{5.27}$$

推出来的组合方差估算公式为，

$$\hat{v}_{6.jrr}(\hat{r}) = (\hat{v}_{4.jrr}(\hat{r}) + \hat{v}_{5.jrr}(\hat{r}))/2。 \tag{5.28}$$

最后，从整个 $2H$ 个虚拟样本中，我们得出，

$$\hat{v}_{7.jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h - \hat{r}_h^C)^2 / 4。 \tag{5.29}$$

类似构造 JRR 方差估算公式的方法，也在 BRR 技术中使用过。对于线性估计

值，偏差校正 JRR 估算公式生成源估算公式，并且，所有的 JRR 方差估算公式都是如此。但对非线性估算公式而言，并非如此。但在实际中，所有 JRR 方差估算公式应当给出非常近似的结果。和 BRR 一样，$\hat{v}_{7.jrr}$可以看成是源估计值 $\hat{\theta}$ 最自然的方差估计值。

JRR 技术可以扩展到更普遍的情形——每个层级抽取两个以上的整群，并使用无放回式来抽取整群。通过连续排除层级中的一个整群，并将保留的整群恰当地给予加权，可以构造虚拟样本和它的补充形式(Wolter,1985;章节 4.6)。和 BRR 一样，我们将比率估计值 $\hat{r}$ 的方差估算 JRR 技术应用于 MFH 调查设计。

范例 5.3

MFH 调查中的 JRR 技术。我们继续讨论 30 ~ 64 岁男性中，CHRON(慢性病)的比率类型子总体比例估计值 $\hat{p}$ 与 SYSBP(收缩血压)子总体均值估计值 $\bar{y}$ 的方差近似估算。使用已有的整群层次的数据，我们计算所有 7 个 $\hat{p}$ 与 $\bar{y}$ 的 JRR 方差估计值。

因为 $H=24$，我们使用弗兰克尔方法构造 24 个 JRR 虚拟样本及它们的补充形式。首先，根据虚拟值 $\hat{p}_h^P$，$\hat{p}_h^{PC}$，$\bar{y}_h^P$ 和 $\bar{y}_h^{PC}$，从虚拟样本和它们的补充形式中计算出式 5.24 与式 5.25 给出的源比率与均值估计值 $\hat{p}$ 与 $\bar{y}$，以及相应的偏差校正估计值。它们是，

$$\hat{p} = 0.3976,\ \bar{\hat{p}}^P = \sum_{k=1}^{24} \hat{p}_k^P/24 = 0.3972 \quad 与 \quad \bar{\hat{p}}^{PC} = \sum_{k=1}^{24} \hat{p}_k^{PC}/24 = 0.3980,$$

$$\bar{y} = 141.785,\ \hat{\bar{y}}^P = \sum_{k=1}^{24} \bar{y}_k^P/24 = 141.793 \quad 与 \quad \hat{\bar{y}}^{PC} = \sum_{k=1}^{24} \bar{y}_k^{PC}/24 = 141.777。$$

CHRON 的 3 个比例估计值以及 SYSBP 的 3 个均值估计值都十分接近。我们接下来计算 JRR 方差估计值。CHRON 比例估计值 $\hat{p}$ 的第一个方差估计值(式 5.20)为，

$$\hat{v}_{1.jrr}(\hat{p}) = \sum_{h=1}^{24} (\hat{p}_h - 0.3976)^2 = 0.1099 \times 10^{-3},$$

使用式 5.21，从补充虚拟样本中，我们得到，

$$\hat{v}_{2.jrr}(\hat{p}) = \sum_{h=1}^{24} (\hat{p}_h^C - 0.3976)^2 = 0.1107 \times 10^{-3}。$$

式 5.22 中的组合方差估计值为，

$$\hat{v}_{3.jrr}(\hat{p}) = (\hat{v}_{1.jrr}(\hat{p}) + \hat{v}_{2.jrr}(\hat{p}))/2 = 0.1103 \times 10^{-3}。$$

使用虚拟值和偏差校正估算公式，可以得到第二组式 5.26 到式 5.29 的 JRR 方差估计值。$\hat{v}_{1.jrr}$的对应估计值是，

$$\hat{v}_{4.jrr}(\hat{p}) = \sum_{h=1}^{24} (\hat{p}_h^P - 0.3972)^2 = 0.1060 \times 10^{-3},$$

从补充虚拟样本中，我们有，

$$\hat{v}_{5.jrr}(\hat{p}) = \sum_{h=1}^{24}(\hat{p}_h^{PC} - 0.3980)^2 = 0.1067 \times 10^{-3}.$$

其中的组合方差估计值为，

$$\hat{v}_{6.jrr}(\hat{p}) = (\hat{v}_{4.jrr}(\hat{p}) + \hat{v}_{5.jrr}(\hat{p}))/2 = 0.1063 \times 10^{-3}。$$

从所有的虚拟样本及它们的补充形式中，我们得到，

$$\hat{v}_{7.jrr}(\hat{p}) = \sum_{h=1}^{24}(\hat{p}_h - \hat{p}_h^c)^2/4 = 0.1103 \times 10^{-3}。$$

正如期望的，CHRON 比例估计值 $\hat{p}$ 的 JRR 方差估计值十分接近。对于 SYSBP 的均值估计值 $\bar{y}$，我们得到以下 JRR 方差估计值，

$$\hat{v}_{1.jrr}(\bar{y}) = \sum_{h=1}^{24}(\bar{y}_h - 141.785)^2 = 0.2773,$$

$$\hat{v}_{2.jrr}(\bar{y}) = \sum_{h=1}^{24}(\bar{y}_h^C - 141.785)^2 = 0.2803,$$

$$\hat{v}_{3.jrr}(\bar{y}) = (\hat{v}_{1.jrr}(\bar{y}) + \hat{v}_{2.jrr}(\bar{y}))/2 = 0.2788,$$

$$\hat{v}_{4.jrr}(\bar{y}) = \sum_{h=1}^{24}(\bar{y}_h^P - 141.793)^2 = 0.2759,$$

$$\hat{v}_{5.jrr}(\bar{y}) = \sum_{h=1}^{24}(\bar{y}_h^{PC} - 141.777)^2 = 0.2789,$$

$$\hat{v}_{6.jrr}(\bar{y}) = (\hat{v}_{4.jrr}(\bar{y}) + \hat{v}_{5.jrr}(\bar{y}))/2 = 0.2774,$$

$$\hat{v}_{7.jrr}(\bar{y}) = \sum_{h=1}^{24}(\bar{y}_h - \bar{y}_h^C)^2/4 = 0.2788。$$

对于 SYSBP，$\bar{y}$ 的 JRR 方差估计值也相当接近。所有比例估计值与均值估计值的 JRR 估算公式均得出了数值上非常相似的结果。所以，实际或是计算的考量都可以成为选择 JRR 方差估算公式的指导。某些复杂调查分析的软件含有折刀技术。

BOOT 技术

与其他样本再使用方法相似，脱靴也可以在复杂抽样设计中用来近似估算非线性估计值的方差。但是，这一方法与 BRR 和 JRR 有许多不同。比如，虚拟样本的生成相去甚远。我们讨论在两级分层等概抽取设计情形下，比率估计值方差估算的脱靴技术。这样的设计中，每个层级中以放回式抽取相同的整群(可以大于2)。我们采用脱靴的简单版本。这种设计由拉奥与吴(Rao and Wu, 1988)当成朴素脱靴引入，并将它称作 BOOT 技术。

假定，从各个 H 个层级中以放回式抽取 $m = a(\geqslant 2)$ 个整群。样本整群的数目为 $m = a \times H$。我们以下面的方式来构造脱靴虚拟样本：

第一步：从层级 h 中的 a 个样本整群中，以放回式抽取一个规模为 a 的简单随机样本。各个层级中的选取相互独立。这 H 个简单随机样本就组成了

一个 m 个整群的脱靴样本。

第二步：重复第一步 K 次，得到总共 K 个独立的脱靴样本。

重要的是，第一步里各个层级中的简单随机样本以放回式抽取，并且层级层次的选取为相互独立。因而，特定的层级中的样本整群可能多次（甚至是 a 次）进入脱靴样本，或是根本一次也没有。

我们讨论比率估计值 $\hat{r}$ 方差估算中 BOOT 技术。脱靴样本 k 的比率估计值表示为 $\hat{r}_k(k=1,\cdots,K)$。脱靴样本估计值 $\hat{r}_k$ 的均值推出脱靴估算公式：

$$\bar{\hat{r}} = \sum_{k=1}^{K} \hat{r}_k / K。\tag{5.30}$$

首先，根据 $\hat{r}_k$ 与脱靴估算公式 5.30 推导出源估计值 $\hat{r}$ 的蒙特卡洛方差估算公式，

$$\hat{v}_{mc}(\hat{r}) = \sum_{k=1}^{K} (\hat{r}_k - \bar{\hat{r}})^2 / K。\tag{5.31}$$

不幸的是，如拉奥与吴（Rao and Wu，1988）所示，这个直观上具有吸引力的估算公式是不可接受的。因为，它对于 $\hat{r}$ 的方差并不是一致的，同时它对于线性估算公式甚至不是无偏的。但是，对于讨论的各个层级抽取同样数目的整群的情形，经过适当换算的蒙特卡洛方差估算公式，是源估计值 $\hat{r}$ 一致的方差估算公式。因而，第一个 BOOT 方差估算公式是，

$$\hat{v}_{1.boot}(\hat{r}) = \frac{a}{a-1}\hat{v}_{mc}(\hat{r}) = \frac{a}{a-1}\sum_{k=1}^{K} (\hat{r}_k - \bar{\hat{r}})^2 / K。\tag{5.32}$$

使用源估计值 $\hat{r}$ 来代替脱靴估计值，得到另一个方差估算公式：

$$\hat{v}_{2.boot}(\hat{r}) = \frac{a}{a-1}\sum_{k=1}^{K} (\hat{r}_k - \hat{r})^2 / K。\tag{5.33}$$

应当注意到，对于朴素脱靴而言，在层级中样本整群数目不同的情形下，没有明显的解决换算问题的方案。拉奥与吴（Rao and Wu，1988）为这样的情形，根据以放回式从层级中抽取规模为 $m_h(\geqslant 1)$ 的简单随机样本，推出了换算脱靴。当选择 m_h 恰当时，有各种脱靴的版本。西特尔（Sitter，1992）根据无放回式再抽样而非放回式再抽样，以及以放回式重复多次，提出了这种方法的扩展形式。拉奥等（Rao et al.，1992）重新定义了换算脱靴，并使之适合于诸如中位值的非光滑函数的方差估算。

在 BOOT 技术中获得足够精度的方差估算结果，脱靴样本的数目 K 应当较大，最好为 500 ~ 1 000。这一技术因而需要较大的处理能力并可能占用许多计算资源。从这一点上，BOOT 技术比 BRR 与 JRR 明显地更依赖计算机。

范例 5.4

MFH 中的 BOOT 技术。我们将 BOOT 技术应用于子总体比例估计值与子总体均值估计值 $\hat{p}$(CHRON) 与 $\bar{y}$(SYSBP) 的方差近似估算。两种估计值均

是比率估计值。MFH 调查的子群由 2 699 个 30 ~ 64 岁男性组成。MFH 设计有 $H = 24$ 个层级，每个层级有 $a = 2$ 个整群，因而每个脱靴样本由 $m = 48$ 个整群组成。在生成脱靴样本时，我们使用整群层次的数据。以放回式独立地从各个层级中抽取两个整群的简单随机样本，我们得到一个脱靴样本。因此，层级中的整群出现在脱靴样本中的次数为 0，1 或 2。而一个层级的样本规模总是 2 个整群。注意，这种样本的数目可以很大，比如，我们有 1 000 个脱靴样本，需要抽取总共 24 000 个规模为 2 的独立样本。在这个例子中，$K = 1\ 000$ 个脱靴样本。

从各个 K 脱靴样本中，计算出模仿源估计值 $\hat{r}$ 的估计值 $\hat{r}_k$。脱靴估计值即是 $\hat{r}_k$ 的平均值。使用 $\hat{r}_k$、脱靴估计值和源估计值，我们最后得出 BOOT 方差估计值 $\hat{v}_{1.boot}(\hat{r})$ 与 $\hat{v}_{2.boot}(\hat{r})$。

给定 1 000 个脱靴样本，图 5.1 给出了脱靴样本 CHRON 与 SYSBP 估计值的分布。CHRON 比例与 SYSBP 均值的源估计值与脱靴估计值（式 5.30）为

$$\hat{p} = 0.397\ 6, \text{脱靴估计值为} \bar{\hat{p}} = 0.397\ 3,$$

$$\bar{y} = 141.785, \text{脱靴估计值为} \hat{\bar{y}} = 141.783。$$

CHRON 比例 $\hat{p}$ 的 BOOT 方差估计值（式 5.32 与式 5.33）分别为，

$$\hat{v}_{1.boot}(\hat{p}) = 2 \times \sum_{k=1}^{1\ 000} (\hat{p}_k - 0.397\ 3)^2 / 1\ 000 = 0.103\ 9 \times 10^{-3}$$

及

$$\hat{v}_{2.boot}(\hat{p}) = 2 \times \sum_{k=1}^{1\ 000} (\hat{p}_k - 0.397\ 6)^2 / 1\ 000 = 0.104\ 0 \times 10^{-3}。$$

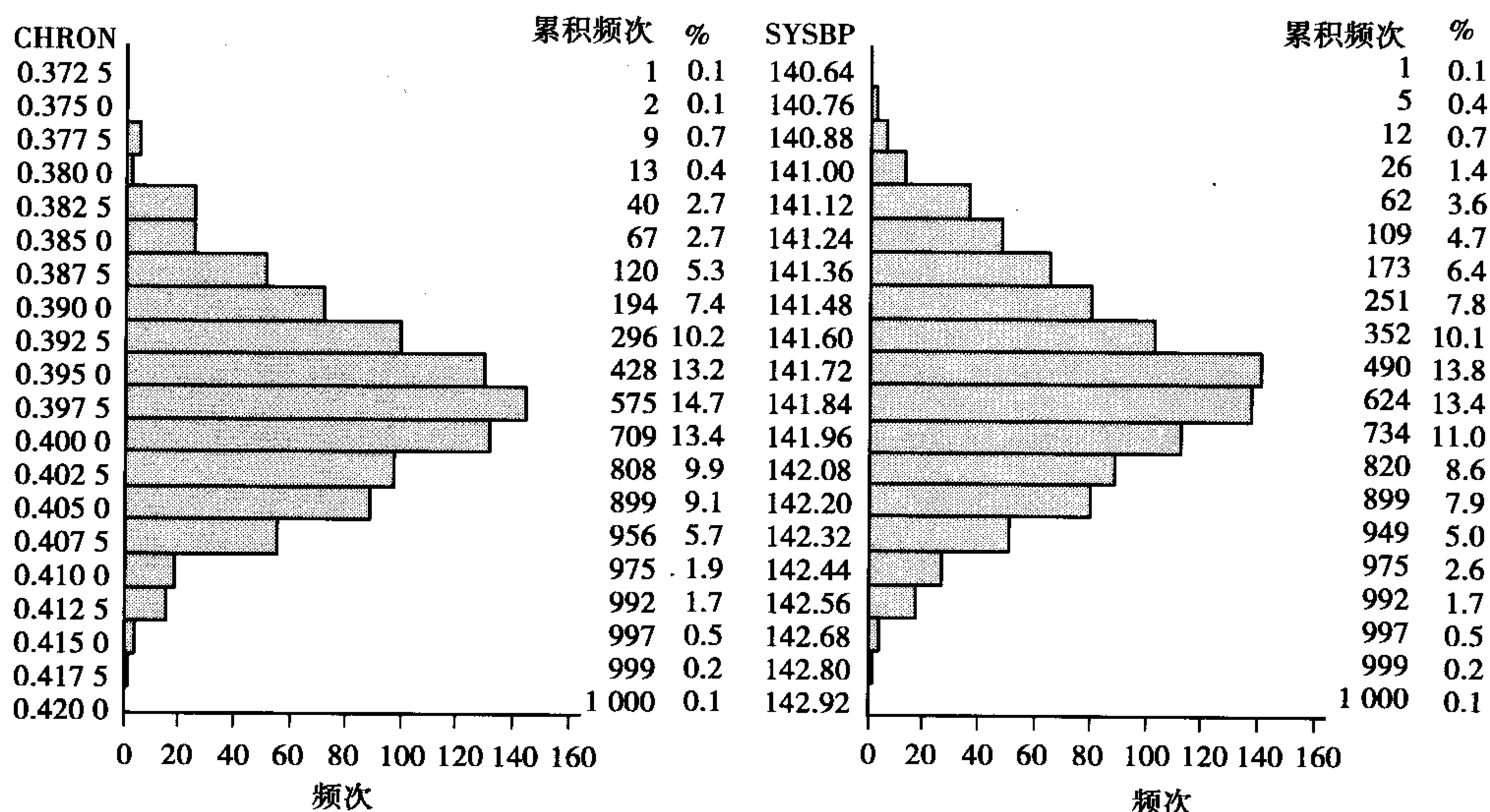


图 5.1 CHRON（二分变量）与 SYSBP（连续变量）脱靴估计值 $\hat{r}_k$ 的脱靴直方图（$K = 1\ 000$ 个脱靴样本）

SYSBP 均值 $\bar{y}$ 的 BOOT 方差估计值为,

$$\hat{v}_{1.boot}(\bar{y}) = 2 \times \sum_{k=1}^{1\,000}(\bar{y}_k - 141.783)^2/1\,000 = 0.279\,8$$

及

$$\hat{v}_{2.boot}(\bar{y}) = 2 \times \sum_{k=1}^{1\,000}(\bar{y}_k - 141.785)^2/1\,000 = 0.279\,8.$$

对于 CHRON 比例估计值 $\hat{p}$ 与 SYSBP 均值估计值 $\bar{y}$,两个 BOOT 方差估计值几乎相等。与其他再使用方法一样,无法对方差估算公式的类型给出任何确定的偏好。从计算的角度,估算公式 $\hat{v}_{2.boot}$ 较 $\hat{v}_{1.boot}$ 简单。

5.5 方差估算公式的比较

线性化方法与样本再使用方法被用作非线性比率估计值方差估算的近似技术。样本假定为从各个层级以放回式抽取至少两个整群的二级等概抽取抽样设计。在层级中整群数目不同(≥2)的设计情形下,我们讨论了线性化方法。平衡半样本(BRR)与折刀重复复制(JRR)技术,要求每个层级中恰好含有两个样本整群,并且层级数目较大。两种方法均被扩展到每个层级中整群数目不同的设计情形。在层级中抽取相等数目(≥2)整群的设计情形下,我们讨论了脱靴技术。同样的,脱靴也被扩展到层级中整群数目不同的设计情形。这些近似方法中,脱靴要求更多的计算资源。接下来,我们比较,以线性化和样本再使用技术从 MFH 调查中得出的方差近似估算的数值结果。

MFH 调查中方差估计值的比较

使用线性化方法——BRR,JRR 与 BOOT 技术,我们估算了二分因变量 CHRON(慢性病)子总体比例估计值,以及连续因变量 SYSBP(收缩血压)子总体均值估计值的方差。它们都是关于 MFH 调查中 2 699 个 30 ~ 64 岁男性子群的比率类型的估计值。详细的结果已在范例 5.1 到范例 5.4 中给出。在 MFH 调查设计中,共有 24 个层级,每个层级含有两个样本整群。这为演示各种方差近似估算法提供了足够的数据。在所有技术中,整群层次的数据有 48 个个案。

表 5.5 给出 CHRON 比例 $\hat{p}$ 与 SYSBP 均值 $\bar{y}$ 的方差与设计效应估计值。设计效应估计值的形式为 deff $= \hat{v}/\hat{v}_{srswr}$,其中,$\hat{v}$ 是所讨论的方差估计值,而 $\hat{v}_{srswr}$ 是对应放回式简单随机抽样的方差估计值。

对于 CHRON 而言,线性化方差估计值、前 3 个 BRR 与 JRR 估计值以及最后的 BRR 与 JRR 估计值都几乎相同。与这些估计值相比,BRR 和 JRR 的第 4,5,6 估计值与两个 BOOT 估计值要小一些。注意,线性化、最后的 BRR

与 JRR 方差估计值(可以看成最恰当的方差估计值)是相同的。对于 SYSBP 而言,所有 BRR 估计值几乎相同,而 JRR 估计值则离异较大。BOOT 方差估计值比其他两组都大些。对于 SYSBP,线性化、最后的 BRR 与 JRR 方差估计值也几乎相同。

表 5.5 MFH 调查 30～64 岁男性 CHRON 比例估计值 $\hat{p}$ 与 SYSBP 均值估计值 $\bar{y}$ 的线性化、BRR、JRR、BOOT 与 SRSWR 方差与设计效应估计值 $\hat{v}$ 及 deff

方 法	慢性病		收缩血压	
	$10^{-3}\times\hat{v}(\hat{p})$	$\text{deff}(\hat{p})$	$\hat{v}(\bar{y})$	$\text{deff}(\bar{y})$
线性化				
DES	0.110 3	1.24	0.278 8	2.06
平衡重复复制				
1	0.110 4	1.24	0.279 1	2.06
2	0.110 3	1.24	0.279 0	2.06
3	0.110 3	1.24	0.279 1	2.06
4	0.105 2	1.18	0.278 7	2.06
5	0.105 6	1.19	0.278 8	2.06
6	0.105 4	1.19	0.278 7	2.06
7	0.110 3	1.24	0.279 0	2.06
折刀重复复制				
1	0.109 9	1.24	0.277 3	2.05
2	0.110 7	1.25	0.280 3	2.07
3	0.110 3	1.24	0.278 8	2.06
4	0.106 0	1.19	0.275 9	2.04
5	0.106 7	1.20	0.278 9	2.06
6	0.106 3	1.20	0.277 4	2.05
7	0.110 3	1.24	0.278 8	2.06
脱靴				
1	0.103 9	1.17	0.279 8	2.07
2	0.104 0	1.17	0.279 8	2.07
SRSWR	0.088 8	1.00	0.135 2	1.00

CHRON 与 SYSBP 的设计效应显示出了不同程度的群内相关。CHRON 的群内相关比 SYSBP 要小很多。对于 SYSBP,各技术间的设计效应估计值仅显示了较小的离异。

作为结论,用线性化方法、BRR、JRR,以及 BOOT 技术获得的两个因变量的比率估计值的方差估计值差别并不大。所以,实际中可以用已有软件或是其他实际要求来选择何种技术。若需要进一步的训练,读者可以借助本书的扩展网页。

方差近似估算法的其他特征

从理论方面,已有的文献根据经验和模拟研究评估了基于线性化、BRR、JRR,以及脱靴的方差近似估算技术。我们简要地指出其中一些结论。

基什与弗兰克尔(Kish and Frankel, 1974)根据一个从每个层级中放回式抽取两个整群的一级分层设计等概抽取样本的经验研究,讨论了线性化、BRR与 JRR 的相对结果。他们指出,首先对于线性估算公式而言,方差估计值相同,并与标准教科书中的估计值一致。但是,非线性估计值——例如比率估计值、回归估计值或是相关系数估计值——的特征则不同。线性化方法给出了最稳定的方差估计值,BRR 的结果最不稳定。但考虑到多种标准时,没有任何一个估计值给出了一致的最好的结果。基什与弗兰克尔得出结论,线性化方法最适合比率估计值,但样本再使用方法则适合其他非线性估计值。

克鲁斯基与拉奥(Krewski and Rao, 1981)展示了线性化方法、BRR 与JRR 有相似的一阶渐进性特征。拉奥与吴(Rao and Wu, 1985)讨论了高阶特征,并显示了线性化方法与 JRR 在每个层级抽取两个整群的放回式设计下,二阶特征也相同。拉奥与吴(Rao and Wu, 1988)讨论了脱靴方法,并展示了其换算的脱靴方差的一阶特征与线性化、BRR、JRR 的相同,但二阶特征不同。与线性化或是 JRR 相比,换算的脱靴估计值的稳定性更差。拉奥等(Rao et al., 1992)讨论了中位值的折刀、BRR 及脱靴方法的方差估计值,没有发现较大的差别。

5.6 职业健康保健调查

在这一小节中,我们描述职业健康保健调查(OHC 调查)已有数据的抽样设计、数据收集及其特征。OHC 调查的抽样设计是一个使用了一级和二级抽样的分层整群抽样的例子。因此,OHC 调查的抽样设计比 MFH 调查要复杂些。同时,OHC 调查中有数量较大的样本整群,这样的设计产生了几个因变量较大的整群效应。所以,这一抽样设计非常适合于在分析复杂调查中考察整群效应。第 7、8 章中的例子也要进一步使用 OHC 调查。

与许多工业化国家一样,芬兰立法管制职业健康服务的提供。1979 年生效的《职业健康服务法案》指导职业健康服务的发展。除了少数的例外,所有雇主被要求为雇员提供职业健康服务,以便他们更关注与工作相关的健康事故。通过《全国疾病保险计划》,雇主从社会保险署返回一定份额的职业健康服务的花销。对于雇员而言,职业健康服务是免费的。为了评估《职业健康服务法案》的运行,实施了多个抽样调查,其中最主要的是 1985 年的调查。

抽样设计

与 MFH 调查相似,OHC 调查也可以归纳为一个多目标的分析性抽样调查。OHC 的目标有,评价 OHC 法案固定行为的执行情况,评估立法的基本目

标达到了什么程度，以及发现职业健康服务可以怎样进一步推进。调查的中心是除去农业和林业的所用工业公司，包括雇主与雇员，以及为这些公司提供OH服务的单位。目标总体含有2百万名雇员和超过100 000个工业公司。

在研究设计中，工业公司是主要的抽样和数据收集单位。因为，芬兰有涵盖目标公司的全国性登记，因而以公司为整群——亦即主要抽样单位(PSU)——的整群抽样成为一个自然的选择。与MFH抽样设计不同，OHC调查中的整群抽样设计的主要原因是主题而非费用效率。

在公司抽样框里，PSU的规模差异很大，从一个人的作坊到上千人的企业。当预计到数据收集的个人层次的样本规模时，需要考虑整群规模差异较大的特征。所以，整群总体根据其规模大小分层，并在规模较大的整群中使用二级抽样。除了规模以外，根据公司的行业划分了6个层级。最多不超过100人的公司的层级使用了一级抽样；其余的使用二级抽样，从这些大的整群中抽取大约50个雇员。这样产生了大约17 000名雇员和1 542个公司的样本。层级层次整群的配额，是根据已知的关于整群平均规模的信息来实施的。这样使得雇员样本接近于等概抽取，即每个雇员的选中概率相等。雷同能(Lehtonen，1988)给出了更详细的抽样设计信息。

数据收集与无应答

从雇主、雇员和OH服务点收集数据时，使用了结构性的问卷。在数据收集过程中，发现有些公司——主要是小公司——已经倒闭了，因而最后完成问卷的公司数目为1 362，应答率为88%。另外，1 195个有应答的公司中，82%(13 355名)的雇员完成了个人问卷。最后，93%的相应的公司OH服务点完成了问卷；这导致了816个OHC涵盖的服务点中，760个给出了相关信息。表5.6给出了最终调查数据中各个层级的公司和雇员数目。

表5.6 OHC调查数据中单位与雇员数目(分层级)

层级	规 模	数 目		平均整群
		单 位	雇 员	样本规模
1	1～10	696	1 730	2.5
2	11～100	176	4 143	23.5
3	101～500	52	2 396	46.1
4	501+	21	976	46.5
5	(所有)	109	1 396	12.8
6	(所有)	141	2 714	19.2
样本总计		1 195	13 355	11.2

部门类型：

层级1～4：除去层级5与6的其他部门；

层级5：建筑部门；

层级6：公共服务部门。

对数模型的分析显示，根据某些结构性因素——规模、行业，以及组织类型，公司问卷的应答率有着统计上的显著差异。图 5.2 给出了相应问卷的估计应答率（根据含有规模、行业、组织类型以及后两者的交互作用的对数模型）。小规模、建筑行业以及仅有一个服务点都增加无应答的概率。

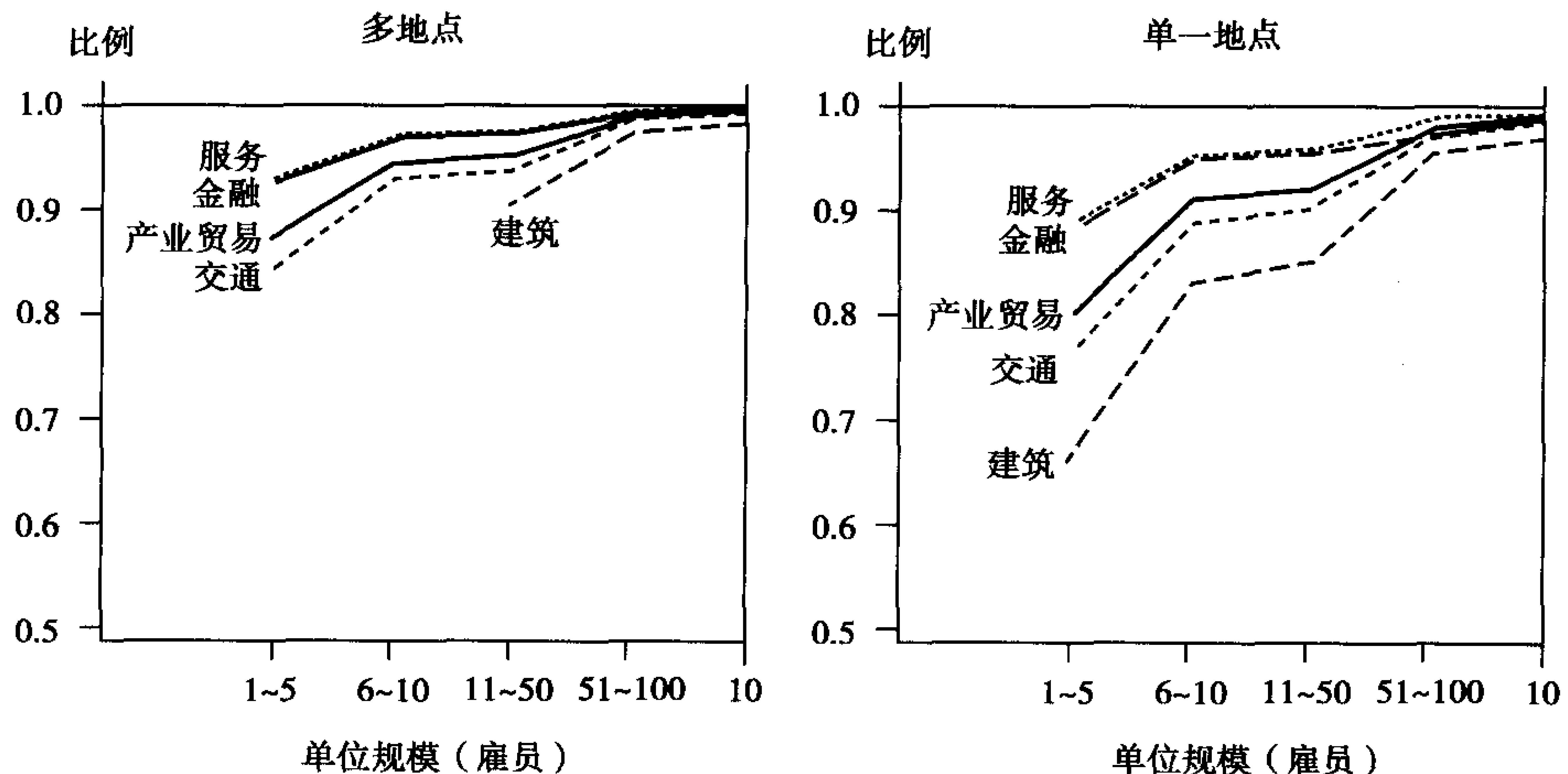


图 5.2　根据对数模型的多地点与单一地点单位的预测应答率（分单位规模与类型）

大公司的无应答相当低，并且无应答与行业及组织类型无关。同时，OH 服务涉及的公司，以及 OHC 法案强制管理的公司，更容易回答相应的问卷。另外，OHC 法案强制管理的公司，不管是否为 OH 服务涵盖，其应答率大致相当。应答率最低的是那些规模最小的、建筑行业的，并且不为 OH 服务涵盖的公司。

在公司层次上的分析——比如对 OHC 涵盖范围的估算，要求对无应答加权。要构造出权重，这样也可以对 PSU 在层级层次不同的选中概率作出补偿。在雇员层次上，抽样设计差不多是等概选取，而无应答公司的雇员总数相对较小。所以，元素层次的分析的无应答校正没有整群层次的重要。比如，有关 OH 服务涵盖的公司在雇员层次上的目标总体的推论中就是如此。

设计效应

在范例中用来演示的 OHC 调查数据是最少有 10 个雇员的公司组成的子群。这个数据包括 5 个层级中的 250 个整群，共有 7 841 个雇员。这一数据可以被看成是近似的自我加权。子群中整群的规模从 10 到 60 个雇员。注意，这个子群是独立组群类型的。表 5.7 给出了含有某些变量的数据。

样本整群——即公司——规模较大（250 个），这有利于协方差矩阵的估算。对某些个体层次的因变量而言，样本中的公司倾向于同质，这导致了群内正相关。比如，在一个制造业公司里，大多数工人的工作条件相似，而这样的

条件与一个办公室公司的不同,而后者内部的同质性也较高。这样生成的均值和比例设计效应的估计值将大于1。特别是对于测量工作场所相关的物理或是社会心理的工作条件的个体层次的变量而言,更是如此。对于一些变量,群内相关要小些,比如,描述总体心灵(心理和精神)压力和心理症状的变量。表5.8给出了一些因变量的设计效应。

表5.7 OHC调查数据,分性别与年龄。慢性病患者(CHRON)、工作中身体健康风险比例及9个心理症状的标准化主成分取值的均值

性别	年龄	样本		CHRON	PHYS	PSYCH
		n	%	%	%	均值
男		4 485	57.2	29.3	46.0	-0.104
女		3 356	42.8	29.2	19.4	0.139
男	15~24	504	6.4	15.5	52.8	-0.300
25~34		1 355	17.3	19.8	50.8	-0.160
35~44		1 453	18.5	27.1	42.9	-0.073
45~54		847	10.8	44.2	41.9	-0.033
55~64		326	4.2	61.3	39.3	0.102
女	15~24	418	5.3	16.0	19.1	0.095
25~34		993	12.7	18.9	18.9	0.132
35~44		1 002	12.8	26.5	17.9	0.104
45~54		681	8.7	43.5	18.5	0.168
55~64		262	3.3	61.8	29.4	0.301
两性	15~24	922	11.8	15.7	37.5	-0.121
25~34		2 348	29.9	19.4	37.4	-0.036
35~44		2 455	31.3	26.9	32.7	-0.000
45~54		1 528	19.5	43.8	31.5	0.056
55~64		588	7.5	61.6	34.9	0.191
样本总计		7 841	100.0	29.2	34.6	0.000

表5.8 OHC调查数据中某些二分变量的比例估计值的平均设计效应估计值(括号里为变量数目)

研究变量	平均 deff
物理工作条件(12)	6.5
社会心理工作条件(11)	3.3
心理症状(8)	2.0
精神症状(9)	1.8

平均设计效应的估计值较大,特别是那些与工作条件紧密相联的变量更是如此。而对于那些不能被认为与工作相关的变量,其均值接近1。我们选择了3个变量来做进一步的分析:二分变量PHYS(身体工作健康事故)与CHRON(慢性病),以及连续变量PSYCH(心灵压力)。PHYS的群内相关很

强，整体设计效应估计值为 7.2。CHRON 与 PSYCH 的整体设计效应估计值分别为 1.8 和 2.0。同时，PHYS 显然是与工作条件相关，而 CHRON 与 PSYCH 则不很明显。

5.7 协方差矩阵估算的线性化方法

加权比率估算

前面，我们讨论了单一的比率估计值。比率估计值向量由 u 个比率估计值组成，其中 $u \geqslant 2$ 是称为组群的总体子群的数目。组群是通过一个或多个定类预测变量的交叉分类形成的。这些变量包括性别、年龄组、社会经济元素或地域变量等。我们的目标是在给定的复杂抽样设计下，一致地估计组群比率参数及其相应的协方差矩阵。为了这一目标，我们为组群比率构造了一个加权比率估计值。对于二分因变量，我们有加权组群比例，而对于连续变量，我们有加权组群均值。

让 N 元素总体分成 u 个互不交叉的子总体或组群。未知总体比率参数是一个列向量，表示为 $\mathbf{R} = (R_1, \cdots, R_u)'$。它由 u 个组群比率参数 $R_j = T_j / N_j$ 组成。其中，T_j 表示因变量的总体组群总和，N_j 表示组群规模，$\sum_{j=1}^{u} N_j = N$。在二分变量例子中，比率参数向量用 $\mathbf{p} = (p_1, \cdots, p_u)'$ 来表示，包含比例参数 $p_j = N_{j1} / N_j$，其中的 N_{j1} 是二分因变量在组群 j 的总体总和。在连续变量例子中，参数向量用 $\overline{\mathbf{Y}} = (\overline{Y}_1, \cdots, \overline{Y}_u)'$ 来表示，其中的 $\overline{Y}_j$ 是组群均值参数 $\overline{Y}_j = T_j / N_j$。用分层整群抽样抽取含有 n 个元素的样本，在各个 $h = 1, \cdots, H$ 层级中抽取 m_h 个整群，样本整群总数为 $m = \sum_{h=1}^{H} m_h$。其中，$H \geqslant 1$，$m \geqslant 2H$，以及 $m > u$。在二级整群抽样中，从层级 h 的整群样本 i 中抽取 n_{hi} 个元素，$\sum_{h=1}^{H} \sum_{i=1}^{m_h} n_{hi} = n$。如果是一级抽样，选取的样本整群中所有元素均进入元素层次的样本中。

在复杂调查中，通常使用总体中元素的选中概率相等的等概抽取设计。这是因为它有利于统计分析。我们在本章前面讨论了这样的设计，MFH 和 OHC 调查抽样设计被当成是等概抽取设计。在实际中，即使是等概抽取设计的各个层级间元素的选中概率也可能不等。因而，校正无应答需要再加权，以便获得一致的估算。为了涵盖这些例子，我们构造了加权比率估计值，它比前面等概抽取样本中讨论过的估计值整体上更适用些。

自加权的数据需要等概抽取设计，个案无应答可以忽略不计。如果数据不是自加权，统计分析要求生成合适的权重变量。权重变量给每个元素赋值，以校正元素不等的选中概率与无应答。基本上如第 2 章所示，样本元素 k 的

权重 $w_k = 1/\pi_k$，即选中概率的倒数。在第 4 章中，引入了权重 $w_k^* = 1/(\pi_k\hat{\theta}_k)$，其中的 $\hat{\theta}_k$ 为估计的应答概率。在非等概抽取设计中，不等的选中概率可能出现，比如因为非比例配额。对于无应答校正，样本数据可以被分成一定数量的校正框，框 c 内的应答率 θ_c 相等，而框间的应答率则可以不等。这些校正框通过利用无应答个案也含有数值的辅助变量来建立。使用后续分层时，通过总体层次的辅助信息（见章节 3.3 与 5.1 以及第 4 章）来构造校正框。注意，在自加权数据中，权重为常数，因为 π_k 与 $\hat{\theta}_k$ 为常数。

如章节 5.1 所示，主要有两种构造权重变量的方法。在估计研究变量的总体总和的描述性调查中，构造权重变量的要求是，n 个元素的权重 w_k^* 的总和 $\hat{N}$ 是总体规模 N 的一致估计值。这种类型的权重在第 2 章到第 4 章中得以广泛使用。在这种总和估计值并不常见的分析性调查中，通常要换算权重，使得其和等于现有的样本数据的规模 n。虽然两种权重变量均在已有的调查分析软件中使用，但是，其和为 n 的换算权重 w_k^{**} 在需要权重变量的统计分析中更有优势。

使用权重 w_k^* 时，构造组合比率估计值向量 $\hat{\mathbf{r}} = (\hat{r}_1, \cdots, \hat{r}_u)'$。它由组群比率估计值 $\hat{r}_j = \hat{t}_j/\hat{N}_j$ 组成，其中，$\hat{t}_j$ 是组群 j 中因变量总体总和 T_j 的加权总和估计值，而 $\hat{N}_j$ 是组群 j 的加权规模，$\sum_{j=1}^{u}\hat{N}_j = \hat{N}$ 是所有 n 样本的权重和。这样，对于相应的总体参数 T_j 与 N_j 而言，t_j 与 $\hat{N}_j$ 是一致的。所以，组群比率估计值 $\hat{r}_j$ 是组群比率 R_j 在给定复杂抽样设计下的一致估计值。

前面组群比率估计值 $\hat{r}_j$ 的加权和 $\hat{t}_j$ 与 $\hat{N}_j$ 得以换算到总体层次。为了分析的目的，我们调整权重，使其和为样本数据规模 n。所以，为了得出 $\hat{r}_j$，我们使用 $\hat{t}_j$ 与 $\hat{N}_j$ 的换算加权形式 y_j 与 x_j，其中 $y_j = (n / \hat{N})\hat{t}_j$，$x_j = (n / \hat{N})\hat{N}_j$，$\sum_{j=1}^{u} x_j = n$。组群比率估计值 $\hat{r}_j$ 可以写成如下形式，

$$\hat{r}_j = \frac{y_j}{x_j} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{m_h} y_{jhi}}{\sum_{h=1}^{H}\sum_{i=1}^{m_h} x_{jhi}} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{k=1}^{x_{hi}} w_{jhik}^{**} y_{jhik}}{\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{k=1}^{x_{hi}} w_{jhik}^{**}}, j = 1, \cdots, u, \tag{5.34}$$

其中，y_{jhi}是层级 h 的整群 i 中组群 j 里因变量元素的加权样本和，x_{jhi}是相应的组群样本规模。式 5.34 中换算过的权重 w_{jhik}^{**} 加总起来为 n。

对于二分变量，元素为式 5.34 形式的比率估计值 $\hat{\mathbf{r}}$ 是用 $\hat{\mathbf{p}} = (\hat{p}_1, \cdots, \hat{p}_u)'$ 来表示的比例估计值。它由组群估计值 $\hat{p}_j = y_j/x_j = \hat{n}_{j1}/\hat{n}_j$ 组成，其中，$\hat{n}_{j1}$ 是组群 j 中的二分变量样本元素的加权样本和，$\hat{n}_j$ 是加权组群规模，$\sum_{j=1}^{u}\hat{n}_j = n$。在

等概抽取设计以及当数据为自加权的情形下,可以获得 $\mathbf{p}$ 的简单未加权估计值 $\hat{\mathbf{p}}^U = (\hat{p}_1^U, \cdots, \hat{p}_u^U)'$。其中,$\hat{p}_j^U = n_{j1}/n_j$ 是组群参数 p_j 的一致估计值,n_{j1} 是组群 j 中的因变量的样本和,n_j 是相应的组群样本规模,而 $\sum_{j=1}^{u} n_j = n$。在这一例子中,$\hat{\mathbf{p}}$ 与 $\hat{\mathbf{p}}^U$ 相同。注意,当数据不是自加权时,估计值 $\hat{\mathbf{p}}^U$ 不是 $\mathbf{p}$ 的一致估计值。

对于连续因变量,我们用 $\bar{\mathbf{y}} = (\bar{y}_1, \cdots, \bar{y}_u)'$表示加权比率估计值向量,组群样本均值 $\bar{y}_j = y_j/x_j$ 是相应的组群总体均值 $\bar{Y}_j = T_j/N_j$ 的一致估计值。相应的未加权的估计值为 $\bar{\mathbf{y}}^U = (\bar{y}_1^U, \cdots, \bar{y}_u^U)'$。

大家可能注意到,实际上比率估计值 $\hat{r}_j$ 所需的数据由 m 个整群层次换算加权和 y_{jhi} 与 x_{jhi} 所组成。而通过使用规模为 m 的整群层次数据来可以达到分析数据的目的,没有必要接触到规模为 n 的元素层次的数据。但在实际中使用调查分析软件时,加权后的样本和 y_j 与 x_j 是通过使用换算过的元素权重 w_{jhik}^{**} 从元素层次数据中估算出来的。

协方差矩阵估算

比率估计值向量 $\hat{\mathbf{p}}$ 的未知总体协方差矩阵 $\mathbf{V}/n$ 含有 u 行和 u 列,所以,它是一个 $u \times u$ 的矩阵。$\mathbf{V}/n$ 是对称的,其下三角与上三角相同。组群比率估计值的方差在 $\mathbf{V}/n$ 主对角线上,而组群比率估计值相应的协方差在矩阵对角线以外。在 $\mathbf{V}/n$ 中共需要估计 $u \times (u+1)/2$ 个不同的参数。

方差和协方差估计值 $\hat{v}_{des}(\hat{r}_j)$ 与 $\hat{v}_{des}(\hat{r}_j, \hat{r}_l)$ 分别占据协方差矩阵估计值 $\hat{\mathbf{V}}_{des}$ 的对角和对角以外的位置。而 $\hat{\mathbf{V}}_{des}$ 是比率估计值向量 $\hat{\mathbf{r}} = (\hat{r}_1, \cdots, \hat{r}_u)'$的渐进性协方差矩阵 $\mathbf{V}/n$ 的一致估计值。方差和协方差估计值从章节 5.3 讨论过的线性化方法推导出来。章节 5.3 给出的单一比率估计值 $\hat{r}$ 的方差与协方差估算公式相应的直接扩展,就是 $\hat{r}_j = y_j/x_j$ 方差估计值 $\hat{v}_{des}(\hat{r}_j)$ 中的样本总和 y_j 与 x_j 的方差与协方差估算公式,以及不同组群中 $\hat{r}_j$ 与 $\hat{r}_l$ 的协方差估计值$\hat{v}_{des}(\hat{r}_j, \hat{r}_l)$ 中的样本总和 y_j, y_l, x_j 与 x_l 的协方差估算公式。因而,我们不写成公式。

与换算的例子一样,$\hat{r}_j$ 与 $\hat{r}_l$ 的方差与协方差估计值是基于放回式假设,而考虑的离异是群间离异。这将引起估计值中的偏差。但如果一级抽样比例很小时,这样的偏差可以忽略不计。

y_j, y_l, x_j 与 x_l 的方差与协方差估计值最后集合在相应的 $u \times u$ 协方差矩阵估计值 $\hat{\mathbf{V}}_{yy}$, $\hat{\mathbf{V}}_{xx}$ 以及 $\hat{\mathbf{V}}_{yx}$ 中。使用这些估计值,通过线性化方法的、基于设计的 $\hat{\mathbf{r}}$ 的协方差矩阵估计值为,

$$\hat{\mathbf{V}}_{des} = \mathrm{diag}(\hat{\mathbf{r}})(\mathbf{Y}^{-1}\hat{\mathbf{V}}_{yy}\mathbf{Y}^{-1} + \mathbf{X}^{-1}\hat{\mathbf{V}}_{xx}\mathbf{X}^{-1} - \mathbf{Y}^{-1}\hat{\mathbf{V}}_{yx}\mathbf{X}^{-1} - \mathbf{X}^{-1}\hat{\mathbf{V}}_{xy}\mathbf{Y}^{-1})\mathrm{diag}(\hat{\mathbf{r}}), \tag{5.35}$$

其中，

$\text{diag}(\hat{\mathbf{r}}) = \text{diag}(\hat{r}_1, \cdots, \hat{r}_u) = \text{diag}(y_1/x_1, \cdots, y_u/x_u)$

$\mathbf{Y} = \text{diag}(\mathbf{y}) = \text{diag}(y_1, \cdots, y_u)$

$\mathbf{X} = \text{diag}(\mathbf{x}) = \text{diag}(x_1, \cdots, x_u)$

$\hat{\mathbf{V}}_{yy}$是样本和 y_j 与 y_l 的协方差矩阵估计值；

$\hat{\mathbf{V}}_{xx}$是样本和 x_j 与 x_l 的协方差矩阵估计值；

$\hat{\mathbf{V}}_{yx}$是和 y_j 与 x_l 的协方差矩阵估计值；

$\hat{\mathbf{V}}_{xy} = \hat{\mathbf{V}}'_{yx}$

运算符号“diag”生成一个对角矩阵，相应的向量元素占据对角线，而对角线以外的则全为0。注意，在线性例子中，协方差矩阵估计值 $\hat{\mathbf{V}}_{xx}$、$\hat{\mathbf{V}}_{yx}$ 以及 $\hat{\mathbf{V}}_{xy}$ 中的元素均为0。

在估算 $\hat{\mathbf{V}}_{des}$ 的元素时，假定至少从 H 个层级中以放回式抽取两个整群。在调查抽样中经常使用的 $m_h=2$ 个整群的特例中，估算公式可以参照章节5.3简化。

作为一个简单例子，让组群数目为 $u=2$。协方差矩阵

$$\hat{\mathbf{V}}_{des} = \begin{bmatrix} \hat{v}_{des}(\hat{r}_1) & \hat{v}_{des}(\hat{r}_1, \hat{r}_2) \\ \hat{v}_{des}(\hat{r}_2, \hat{r}_1) & \hat{v}_{des}(\hat{r}_2) \end{bmatrix}$$

其中的元素为：

方差估计值，

$$\hat{v}_{des}(\hat{r}_j) = \hat{r}_j^2[y_j^{-2}\hat{v}(y_j) + x_j^{-2}\hat{v}(x_j) - 2(y_jx_j)^{-1}\hat{v}(y_j, x_j)], j = 1,2。$$

协方差估计值，

$$\hat{v}_{des}(\hat{r}_1, \hat{r}_2) = \hat{r}_1\hat{r}_2[(y_1y_2)^{-1}\hat{v}(y_1, y_2) + (x_1x_2)^{-1}\hat{v}(x_1, x_2) - (y_1x_2)^{-1}\hat{v}(y_1, x_2) - (y_2x_1)^{-1}\hat{v}(y_2, x_1)]。$$

由于 $\hat{\mathbf{V}}_{des}$ 的对称性，估计值 $\hat{v}_{des}(\hat{r}_1, \hat{r}_2)$ 与 $\hat{v}_{des}(\hat{r}_2, \hat{r}_1)$ 相等。当估计值 $\hat{r}_j$ 被当成信息估计值时，分母 x_j 被假定为固定。这种情形下，方差与协方差估计值 $\hat{v}(x_j)$ 与 $\hat{v}(y_j, x_j)$ 为0，$\hat{v}_{des}(\hat{r}_j) = \hat{v}(y_j)/x_j^2$。在二分因变量的情形下，这一估算公式简化成 $\hat{v}_{bin}(\hat{p}_j) = \hat{p}_j(1-\hat{p}_j)/n_j$。

应当注意到，$\hat{\mathbf{V}}_{des}$ 与分布无关，所以它并不需要假定具体的样本观测值的分布。这样，$\hat{\mathbf{V}}_{des}$ 的估计值可以是非对角形的。$\hat{\mathbf{V}}_{des}$ 的非对角形是因为不同组群的比率估计值 $\hat{r}_j$ 与 $\hat{r}_l$ 的相关系数可以不为0。相反，本小节讨论的二分协方差矩阵估计值，由定义其值为0。

不同组群中估计值 $\hat{r}_j$ 与 $\hat{r}_l$ 的非0相关系数的一个来源是样本的整群分类。取决于组群的类型，可以发现各个程度的相关。当组群平滑地切割样本

整群时，样本组群中不同的元素可能分别属于 j 与 l 组群中，例如跨类型的人口或相关因素。如果整群分类的效应明显，则可能发现较大的相关系数。相反，当组群完全分离，一个给定的样本整群就是一个组群时，则得到估计值 $\hat{r}_j$ 与 $\hat{r}_l$ 间的 0 相关系数。例如，当家庭户为一个整群时，典型的整群类因素是家庭净收入与家庭规模，而家庭成员的年龄和性别则是个人类因素。实际中经常遇见的混合型组群则位于中间，非 0 相关占据一些维度，而 0 相关则占据另一些。

发现不稳定性

在给定复杂抽样设计的情形下，式 5.35 中的协方差矩阵估计值是渐进性协方差矩阵 $\mathbf{V}/n$ 的一致估计值。因而，当整群规模固定时，增加样本整群数目 m 可以促成估计值向 $\mathbf{V}/n$ 的一致估计值的收敛。但，当 m 较小时，$\hat{\mathbf{V}}_{des}$ 可能不稳定，即接近于奇异。当组群数目 u 较大时，这也可能发生，因为这需要估算数百个方差与协方差项。协方差矩阵估计值的不稳定性在构造逆矩阵时将造成诸多问题，这将严重影响到检验与建模过程中的可靠性。

当估算渐进性协方差矩阵 $\mathbf{V}/n$ 的自由度 f 较小时，将出现近似奇异或是不稳定性问题。对于标准的复杂抽样设计，f 可以当成样本整群数目中除去层级数，即 $f = m - H$。当 f 相对于组群数目 u，或者更具体的，相对于拟合模型的残差自由度较大时，可以得到稳定的 $\hat{\mathbf{V}}_{des}$。在实际中，当样本整群数目较大，并且 u 远远小于 m 时，可能并不会遇到不稳定的问题。

统计条件数可以用作 $\hat{\mathbf{V}}_{des}$ 不稳定性的测量值。定义比率 $\mathrm{cond}(\hat{\mathbf{V}}_{des}) = \hat{\lambda}_{max}/\hat{\lambda}_{min}$，其中 $\hat{\lambda}_{max}$ 与 $\hat{\lambda}_{min}$ 分别是 $\hat{\mathbf{V}}_{des}$ 最大与最小特征值。当这一统计量较大时，如上百或是上千，则有不稳定性问题。当它较小，如小于 50 时，就没有严重的不稳定性问题。不幸的是，这一统计量并不是调查分析软件的常规输出。下面的表格给出了 MFH 与 OHC 调查中二分因变量 CHRON(慢性病)的比例估计值向量在各个 u 数值下 $\hat{\mathbf{V}}_{des}$ 的条件数。各个调查中的组群由应答者性别和等规模的年龄组构成。

组群数目	MFH	OHC
4	6.5	2.8
8	10.6	3.5
12	39.8	3.6
20	421.5	5.6
24	423 684	6.6
40	n. a.	9.9

n. a.：不适用。

注意,MFH 调查中,$f=24$;OHC 调查中,$f=245$。所以,在 MFH 调查中,可能的最大 u 值为 24,而相应的,$\hat{\mathbf{V}}_{des}$变得很不稳定。而 u 值小于 12 时,估计值相当稳定。在 OHC 调查中,条件数随着 u 的增加略微变大,但 $\hat{\mathbf{V}}_{des}$在 u 的所有取值上均显出了稳定性。协方差矩阵估计值 $\hat{\mathbf{V}}_{des}$的这些特征也可以用图来表示。图 5.3 显示了,MFH 与 OHC 调查设计中,在 $u=24$ 组群的情况下的 CHRON 的估计值 $\hat{\mathbf{V}}_{des}$。对于 MFH 调查,$\hat{\mathbf{V}}_{des}$的不稳定性由矩阵对角线以外的高"尖点"所表示。而 OHC 调查设计中 $\hat{\mathbf{V}}_{des}$的稳定性也清晰可见。

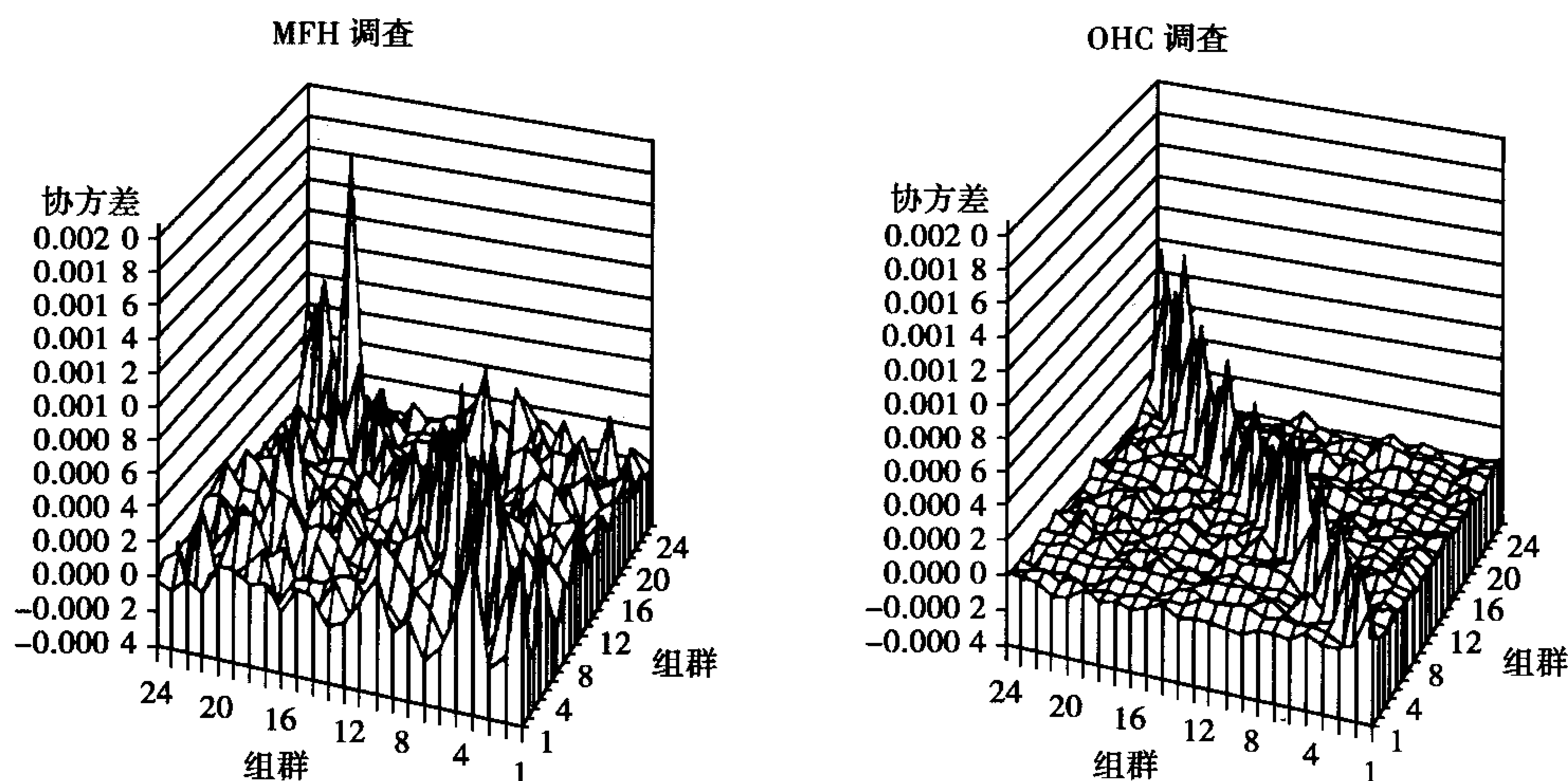


图 5.3 MFH 与 OHC 调查设计中 CHRON 在 $u=24$ 组群比例估计值的协方差矩阵估计值 $\hat{\mathbf{V}}_{des}$

设计效应矩阵估计值

对于设计效应矩阵估计值,我们推导比例估计值向量的二项协方差矩阵估计值。使用二项和相应的基于设计的协方差矩阵估计值,可以得到设计效应矩阵。从设计效应矩阵的对角线抽出的设计效应估计值,可以用来推导说明额外二项离异的协方差矩阵估计值。

为了构造设计效应矩阵估计值,我们不仅需要比例向量基于设计的协方差矩阵估计值,还需要二项变量的相应估计值。对于二分因变量,我们假定一个比例向量 $\hat{\mathbf{p}}$ 的二项抽样模型。假定组群 j 中的加权成功数目由二项分布生成,而这一生成过程在 u 个组群间假定为是相互独立的。比例估计值 $\hat{\mathbf{p}}$ 的协方差矩阵估计值 $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}})$ 是一个对角元素从二项分布中得出的对角矩阵,如下,

$$\hat{v}_{bin}(\hat{p}_j) = \hat{p}_j(1-\hat{p}_j)/\hat{n}_j,\ j=1,\cdots,u。 \tag{5.36}$$

对于未加权比例向量 $\hat{\mathbf{p}}^{U}$，用 $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}}^{U})$ 表示的相应的估计值是通过使用(换算过的)等于 1 的元素权重获得的。应当强调，在式 5.36 的二项方差估计值的分母中，使用了加权观测数 $\hat{n}_j$，即第 j 组群的期望样本规模。观测到的组群样本规模 n_j 可在分母中替代其期望值。

使用基于设计的协方差矩阵估计值 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}})$ 与其二项相对值 $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}})$，可以推导出相应的组群比例估计值向量 $\hat{p}$ 的设计效应矩阵估计值如下，

$$\hat{\mathbf{D}} = \hat{\mathbf{V}}_{bin}^{-1}(\hat{\mathbf{p}})\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}), \tag{5.37}$$

其中，$\hat{\mathbf{V}}_{bin}^{-1}$ 是 $\hat{\mathbf{V}}_{bin}$ 的倒数。设计效应估计值 $\hat{d}_j$ 与 $\hat{p}_j$ 是设计效应矩阵估计值(因而称为设计效应矩阵)的对角元素。设计效应矩阵的特征值 $\hat{\delta}_j$ 通常被称为通用设计效应。设计效应估计值的和等于特征值的和。这一数值可以从 $\hat{\mathbf{D}}$ 对角元素的和——即，其秩——中获得。只有在 $\hat{\mathbf{V}}_{des}$ 也是对角形的特殊例子中，设计效应估计值与相应的特征值才相等。在式 5.37 中的第一个协方差矩阵估计值为对角矩阵，如 $\hat{\mathbf{V}}_{bin}$ 时，所有这些都将成立。但在这些都不成立的、比例更为复杂的情形下，设计效应并不是 $\hat{\mathbf{D}}$ 的对角元素，其和也不等于特征值之和。这些更复杂的设计效应矩阵有时被称为通用设计效应矩阵。第 7 章与第 8 章将讨论它。

比例估计值 $\hat{p}_j$ 的设计效应估计值的形式为，

$$\hat{d}_j = \hat{v}_{des}(\hat{p}_j)/\hat{v}_{bin}(\hat{p}_j), \ j = 1, \cdots, u, \tag{5.38}$$

其中的方差估计值 $\hat{v}_{des}$ 是 $\hat{\mathbf{V}}_{des}$ 的对角元素。设计效应估计值 $\hat{d}_j$ 测量由整群分类引起的比例估计值 $\hat{p}_j$ 中的额外二项离异。当设计效应估计值大于 1 时，就有额外二项离异。

在式 5.38 的二项方差估计值中，如果使用了观测到的组群样本规模，而非期望值，特别是当组群样本规模的期望与观测规模 n_j 与 $\hat{n}_j$ 相去甚远时(这种情形可能在非比例样本配额中出现)，则可以获得不同的设计效应估计值。因而，使用某种软件计算出来的子群比例估计值的设计效应估计值，可能与使用另外软件的结果不同。显然，在自加权样本中，两种方法应当得出相等的设计效应估计值。

应当注意到，在式 5.37 的设计效应矩阵估计值中，仅纳入了整群分类的作用份额。这是因为，它使用了一致加权比例向量估计值的二项协方差矩阵估计值。如果在式 5.37 中使用未加权的比例向量估计值的二项协方差矩阵估计值，而非加过权的比例向量估计值，则所有复杂抽样对于协方差矩阵估算的作用——如不等选中概率、整群分类，以及无应答校正——均得以反映。显然，对于自加权样本，两种方法将得出相等的设计效应估计值。如果使用一致

比例估计值 $\hat{\mathbf{p}}$ 成为原则，则使用加权观测值以及式 5.37 是合理的。因此，要强调校正整群分类在分析复杂调查中的决定性角色。但是，使用两种版本的二项协方差矩阵估计值来计算 deff 矩阵估计值，在评估加权对于设计效应的作用非常有用。

范例 5.5

协方差矩阵和设计效应矩阵的线性化方法估算。使用 OHC 调查数据，我们将详细地计算在由性别变量形成的 $u=2$ 的简单情形下，二分变量 PHYS（身体健康工作事故）的比例估计值 $\hat{\mathbf{p}}$ 及连续变量 PSYCH（9 个心理症状中的第一个标准化主成分）的均值估计值 $\bar{\mathbf{y}}$ 的协方差矩阵估计值 $\hat{\mathbf{V}}_{des}$。因此，$\hat{\mathbf{V}}_{des}$ 是一个 2×2 的矩阵，而组群则是跨层级类型。表 5.9 给出了协方差矩阵估算所需的部分数据。注意，这些数据是整群层次的，由 5 个层级中的 $m=250$ 个整群组成。因此，自由度 $f=245$，雇员层次的样本规模为 $n=7\ 841$。

表 5.9　两个性别形成的组群的 $i=1,\cdots,250$ 个整群中整群层次回应变量 PHYS 与 PSYCH 的样本和 y_{1i}（男）与 y_{2i}（女）及相应的整群样本规模 x_{1i}（男）与 x_{2i}（女）（OHC 调查）

层级	整群	PHYS		PSYCH			
h	i	y_{1i}	y_{2i}	y_{1i}	y_{2i}	x_{1i}	x_{2i}
2	1	11	3	-0.143 4	-0.032 2	36	22
2	2	18	4	-0.192 5	0.186 7	57	21
2	3	4	5	0.004 5	0.367 4	9	15
2	4	2	2	0.713 5	-0.367 9	12	15
2	5	1	0	-0.168 1	0.123 5	27	8
2	6	1	0	-0.267 3	0.150 4	19	21
2	7	9	4	0.009 9	0.209 9	23	27
2	8	4	2	0.368 1	0.015 5	16	31
2	9	0	0	-0.503 3	0.075 5	6	6
2	10	3	0	-0.317 6	-0.251 6	8	8
2	11	2	7	0.974 6	0.190 3	6	67
2	12	7	3	-0.336 1	0.557 2	22	31
2	13	4	1	-0.2329	-0.218 1	9	7
2	14	0	0	-0.203 2	0.589 3	13	16
2	15	1	23	0.413 7	0.256 5	4	56
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
6	245	14	2	0.198 4	-0.427 1	23	7
6	246	2	1	-0.104 9	0.390 5	7	7
6	247	4	7	-0.296 1	0.501 8	7	13
6	248	0	1	-0.807 3	0.927 8	3	9
6	249	2	0	0.000 6	-0.348 4	16	13
6	250	13	1	-0.127 3	-0.146 6	26	4
样本总计		2 061	650	-26.750 1	33.798 3	4 485	3 356

比率估计值 $\hat{\mathbf{r}} = (\hat{r}_j, \hat{r}_l)' = (y_1/x_1, y_2/x_2)'$，其中 $\hat{r}_1$ 与 $\hat{r}_2$ 如式 5.34 中所示。对于二分变量 PHYS，我们用 $\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2)'$ 来表示比率估计值，而连续变量 PSYCH 用 $\bar{\mathbf{y}} = (\bar{y}_1, \bar{y}_2)'$ 表示。以下的 PHYS 数字从表 5.9 中计算得来。

整群层次样本和 y_{jhi} 与 x_{jhi} 的和：

$$\hat{n}_{11} = y_1 = 2\,061 \quad 与 \quad \hat{n}_1 = x_1 = 4\,485(男性),$$

$$\hat{n}_{21} = y_2 = 650 \quad 与 \quad \hat{n}_2 = x_2 = 3\,356(女性)。$$

PHYS 的比例估计值，即 $\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2)'$ 的元素：

$$\hat{p}_1 = y_1/x_1 = \frac{2\,061}{4\,485} = 0.459\,5(男性),$$

以及

$$\hat{p}_2 = y_2/x_2 = 650/3\,356 = 0.193\,7(女性)。$$

我们接下来为计算 PHYS 比例估计值 $\hat{\mathbf{p}}$ 的协方差估计值 $\hat{\mathbf{V}}_{des}$ 构造 2×2 对角矩阵 $\mathrm{diag}(\hat{\mathbf{p}})$，$\mathbf{Y}$ 以及 $\mathbf{X}$。

$$\mathrm{diag}(\hat{\mathbf{p}}) = \begin{bmatrix} 0.459\,5 & 0 \\ 0 & 0.193\,7 \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} 2\,061 & 0 \\ 0 & 650 \end{bmatrix}$$

及

$$\mathbf{X} = \begin{bmatrix} 4\,485 & 0 \\ 0 & 3\,356 \end{bmatrix}。$$

也是从表 5.9 中的整群层次数据得出的协方差矩阵估计值 $\hat{\mathbf{V}}_{yy}$，$\hat{\mathbf{V}}_{xx}$ 以及 $\hat{\mathbf{V}}_{yx}$ 如下：

$$\hat{\mathbf{V}}_{yy} = \begin{bmatrix} 15\,722.50 & -130.45 \\ -130.45 & 3\,261.71 \end{bmatrix},$$

$$\hat{\mathbf{V}}_{xx} = \begin{bmatrix} 34\,560.23 & -7\,315.43 \\ -7\,315.43 & 34\,099.04 \end{bmatrix},$$

及

$$\hat{\mathbf{V}}_{yx} = \begin{bmatrix} 18\,973.88 & -5\,907.69 \\ -1\,098.11 & 6\,051.14 \end{bmatrix} = \hat{\mathbf{V}}'_{xy}。$$

使用这些矩阵，我们最后计算 PHYS 比例的、由式 5.35 给出的协方差矩阵估计值 $\hat{\mathbf{V}}_{des}$。因而，我们有，

$$\hat{\mathbf{V}}_{des} = \begin{bmatrix} \hat{v}_{des}(\hat{p}_1) & \hat{v}_{des}(\hat{p}_1, \hat{p}_2) \\ \hat{v}_{des}(\hat{p}_2, \hat{p}_1) & \hat{v}_{des}(\hat{p}_2) \end{bmatrix} = 10^{-4}\begin{bmatrix} 2.775 & 0.576 \\ 0.576 & 1.951 \end{bmatrix}。$$

举例，使用计算得来的估计值，得出的方差估计值 $\hat{v}_{des}(\hat{p}_1)$ 为，

$$\hat{v}_{des}(\hat{p}_1) = 0.459\,5^2 \times [2\,061^{-2} \times 15\,722.50 + 4\,485^{-2} \times 34\,560.23 - 2 \times (2\,061 \times 4\,485)^{-1} \times 18\,973.88] = 0.277\,5 \times 10^{-3}。$$

$\hat{p}_1$ 与 $\hat{p}_2$ 之间的相关系数为较大的 0.25，显示组群跨越整群边界。$\hat{\mathbf{V}}_{des}$ 的条件数为 $\text{cond}(\hat{\mathbf{V}}_{des})=1.9$，显示了大数 f 与小数 u 引起的稳定性。

对于 PSYCH，以下数字从表 5.9 中计算得来。

整群层次样本和 y_{jhi} 与 x_{jhi} 的和：

$$y_1=-26.7501 \quad 与 \quad x_1=4485(男性),$$
$$y_2=33.7983 \quad 与 \quad x_2=3356(女性)。$$

PSYCH 的均值估计值，即 $\bar{\mathbf{y}}=(\bar{y}_1,\bar{y}_2)'$ 中的元素：

$$\bar{y}_1=y_1/x_1=-0.1008(男性),$$

以及

$$\bar{y}_2=y_2/x_2=0.1347(女性)。$$

2×2 对角矩阵 $\text{diag}(\bar{\mathbf{y}})$，$\mathbf{Y}$ 和 $\mathbf{X}$ 的构造与 PHYS 同出一辙。协方差矩阵估计值 $\hat{\mathbf{V}}_{xx}$ 与 PHYS 的相同，而协方差矩阵估计值 $\hat{\mathbf{V}}_{yy}$ 与 $\hat{\mathbf{V}}_{yx}$ 为：

$$\hat{\mathbf{V}}_{yy}=\begin{bmatrix}6765.34 & 1036.34\\ 1036.34 & 6585.20\end{bmatrix},$$

$$\hat{\mathbf{V}}_{yx}=\begin{bmatrix}-3139.98 & 2129.01\\ -2051.46 & 2259.73\end{bmatrix}=\hat{\mathbf{V}}'_{xy}。$$

使用这些矩阵，我们计算 PSYCH 均值的协方差矩阵估计值 $\hat{\mathbf{V}}_{des}$。

$$\hat{\mathbf{V}}_{des}=\begin{bmatrix}\hat{v}_{des}(\bar{y}_1) & \hat{v}_{des}(\bar{y}_1,\bar{y}_2)\\ \hat{v}_{des}(\bar{y}_2,\bar{y}_1) & \hat{v}_{des}(\bar{y}_2)\end{bmatrix}=10^{-4}\begin{bmatrix}3.223 & 0.427\\ 0.427 & 5.856\end{bmatrix}。$$

PHYS 比例与 PSYCH 均值基于设计的协方差矩阵估计值结果，包括它们的标准误估计值 $\text{s.e}_{des}(\hat{r}_j)$ 如下所示，

j	组群	PHYS		PSYCH		$\hat{n}_j$
		$\hat{p}_j$	$\text{s.e}_{des}(\hat{p}_j)$	$\bar{y}_j$	$\text{s.e}_{des}(\bar{y}_j)$	
1	男	0.460	0.0167	-0.1008	0.0180	4485
2	女	0.194	0.0140	0.1347	0.0242	3356
	样本总和	0.346	0.0144	0.0000	0.0158	7841

使用合适的校正分析软件，可以从表 5.9 中给出的整群层次数据计算出方差和协方差估计值 $\hat{\mathbf{V}}_{yy}$，$\hat{\mathbf{V}}_{xx}$ 与 $\hat{\mathbf{V}}_{yx}$。$\hat{\mathbf{V}}_{des}$ 公式中的矩阵运算，可以由任何合适的矩阵代数软件来执行。但在实际中，应用合适的调查分析软件于元素层次数据来估算 $\hat{\mathbf{V}}_{des}$ 更为方便。总体而言，由几个定类变量构成的 u 个组群的情形下，线性 ANOVA 模型可以用来拟合有着恰当抽样设计选项的、含有因变量的完全交互作用的、不含截距的模型。模型系数则等于组群的比例或均值估

计值,而模型系数的协方差矩阵估计值则提供比例或均值的协方差矩阵估计值 $\hat{\mathbf{V}}_{des}$。

我们接下来计算设计效应矩阵。为了这一目标,需要二项协方差矩阵估计值。

对于 PHYS,计算比例向量 $\hat{\mathbf{p}}$ 二项协方差矩阵估计值中的元素,

$$\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}}) = \begin{bmatrix} \hat{v}_{bin}(\hat{p}_1) & 0 \\ 0 & \hat{v}_{bin}(\hat{p}_2) \end{bmatrix} = \begin{bmatrix} \hat{p}_1(1-\hat{p}_1)/\hat{n}_1 & 0 \\ 0 & \hat{p}_2(1-\hat{p}_2)/\hat{n}_2 \end{bmatrix}$$

我们有,

$$\hat{p}_1(1-\hat{p}_1)/\hat{n}_1 = 0.459\,5(1-0.459\,5)/4\,485 = 0.000\,055\,4(\text{男性}),$$

及

$$\hat{p}_2(1-\hat{p}_2)/\hat{n}_2 = 0.193\,7(1-0.193\,7)/3\,356 = 0.000\,046\,5\ (\text{女性})。$$

将这些方差估计值带入 $\hat{\mathbf{V}}_{bin}$,我们有,

$$\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}}) = 10^{-4}\begin{bmatrix} 0.554 & 0 \\ 0 & 0.465 \end{bmatrix}。$$

应当注意到,协方差矩阵估计值 $\hat{\mathbf{V}}_{bin}$是对角形,因为假定比例估计值 $\hat{p}_1$ 与 $\hat{p}_2$ 不相关。$\hat{\mathbf{V}}_{bin}$的估计值,即使是方差估计值,没有考虑整群分类的效应。所以,当群内相关为正时,二项方差估计值 $\hat{v}_{bin}(\hat{p}_j)$倾向于低估相应的方差。当计算估计值 $\hat{\mathbf{p}}$ 的设计效应矩阵估计值 $\hat{\mathbf{D}} = \hat{\mathbf{V}}_{bin}^{-1}\hat{\mathbf{V}}_{des}$时,就出现了这种情形:

$$\begin{aligned}\hat{\mathbf{D}}(\hat{\mathbf{p}}) &= \begin{bmatrix} 18\,058.295 & 0 \\ 0 & 21\,489.421 \end{bmatrix} \times 10^{-4}\begin{bmatrix} 2.775 & 0.576 \\ 0.576 & 1.951 \end{bmatrix} \\ &= \begin{bmatrix} 5.01 & 1.04 \\ 1.24 & 4.19 \end{bmatrix}。\end{aligned}$$

$\hat{\mathbf{D}}$ 的对角线上的设计效应估计值 $\hat{d}_j$ 为,

$$\hat{d}(\hat{p}_1) = \hat{v}_{des}(\hat{p}_1)/\hat{v}_{bin}(\hat{p}_1) = 0.000\,277\,5/0.000\,055\,4 = 5.01\ (\text{男性})$$

及

$$\hat{d}(\hat{p}_2) = \hat{v}_{des}(\hat{p}_2)/\hat{v}_{bin}(\hat{p}_2) = 0.000\,195\,1/0.000\,046\,5 = 4.19(\text{女性})。$$

这些估计值相当大,显示了因变量 PHYS 的较强整群分类效应。当使用二项协方差矩阵估计值 $\hat{\mathbf{V}}_{bin}$时,这一结果严重低估了估计值 $\hat{p}_j$ 的标准误。除了设计效应估计值以外,还可以计算设计效应矩阵的特征值,即通用设计效应。它们是 $\hat{\delta}_1 = 5.81$,$\hat{\delta}_2 = 3.39$。可以注意到,设计效应估计值的和为 9.20,它等于特征值的和。设计效应估计值的均值为 4.60,显示出了性别的较强的平均整群分类效应。但是,这一均值显著小于从整个样本计算出来的比例估计值 $\hat{p}$ 的整体设计效应估计值 $\hat{d} = 7.2$。这是由设计效应估计值的特征引起的。与

整体设计效应估计值相比，这些设计效应估计值在跨越组群边界的类型中变得较小。

PHYS 均值的估计值如下：

j	组群	$\hat{p}_j$	s. e$_{des}$	s. e$_{bin}$	$\hat{d}_j$	$\hat{n}_j$
1	男	0.460	0.016 7	0.007 4	5.01	4 485
2	女	0.194	0.014 0	0.006 8	4.19	3 356
	样本总计	0.346	0.014 4	0.005 4	7.17	7 841

5.8 本章小结与更多的文献

小　结

在复杂调查中，恰当估算比率估计值非常重要。首先，需要方差估计值来得出比率估计值类的非线性估计值的标准误和置信区间。在等概抽取的二级分层整群抽样设计的情形下，我们估算了比率均值和比率比例估计值的方差。这种情形下的样本数据被假定为自加权的，因而无应答校正是不必要的。从小型芬兰健康调查(MFH 调查)中修改抽样设计得来的演示数据满足了这些条件。

在样本子群的规模 x 并不由抽样设计所固定的重要例子中，我们考察了子群均值与比例的比率类型估计值 $\hat{r} = y/x$。因此，$\hat{r}$ 中分母 x 是一个随机变量，涉及本身的方差以及与分子 y 的协方差。在 y 的方差之外，这些方差与协方差项都成为以线性化方法计算的比率估计值方差估计值的一部分。由于它在实际中的广泛应用及其在调查分析软件中的流行，我们详细地讨论了这一方法。

根据样本再使用方法，我们还讨论了比率估计值方差估算的其他方法。平衡半样本(BRR)与折刀(JRR)技术是传统的样本再使用方法，而脱靴(BOOT)仅在最近才应用于复杂调查之中。与线性化方法不同，这些方法需要较多的计算。但即使这样，它们业已成熟，并可应用于不同类型的非线性估计值。所有近似估算方法都假定了放回式整群抽样。使用这样的假定，仅仅用群间离异就可以评估比率估计值的离异性。相对于根据放回式简单随机抽样的方差估计值，我们使用了设计效应来广泛地测量整群分类对于方差估计值的影响作用。

在方差估算中，选取 MFH 调查抽样设计的原因是它的简单：在修改后的抽样设计中，每个层级中正好含有两个整群。在所有方差近似估算方法中，均使用了涵盖 30～64 岁男性的 MFH 调查数据的子群。选取这一具体子群而非

整个 MFH 调查样本的原因,是总体样本规模是固定的,而讨论的子总体的样本规模则是一个随机变量。这成为使用近似估算方法来演示方差估算的较好对象。选取的子群由恰当地模拟了 MFH 调查数据抽样设计基本特征的跨越整群边界的组群组成。例如,子群中包含了所有 24 个层级 48 个整群的元素。如果仅选取一个地域样本,仅有部分层级和样本整群包含在其中的话,情况就有所不同。

各种方差近似估算方法在二分因变量 CHRON(慢性病;较小群内相关)的比例估计值与连续变量 SYSBP(收缩血压;更强的群内相关)的均值估计值的估算中,得出了相似的结果。由于没有理论来选择何种方差近似估算方法,类似可用软件等技术因素就成了实际中选用方法的指南。

通过在组合比率估计值中使用恰当的各个子群的元素权重,可以估算集结在比率向量中的几个子群比率。在非等概抽取复杂抽样设计的情形下,这生成了比率的一致估算。使用线性化方法,可以得出加权子群比率估计值向量的协方差矩阵的一致估算。我们演示了,因变量群内正相关不仅增大方差估计值,而且还引入不同子群间比率估计值的非 0 相关系数;渐进性有效的协方差估计值的推出,消除了额外的离异和非 0 相关。估计值基本上是非对角的,协方差在对角线外,并且更多地出现在跨越组群边界的情形下。对数与线性模型中,渐进性建模过程需要这样的协方差矩阵估计值。

线性化方法计算出来的协方差矩阵估计值在样本整群数目较小的情形下可能会不稳定。在标准误估算以及检验和建模过程中,不稳定将造成问题。根据统计条件数或是协方差矩阵估计值的图形检查,有技术可以发现不稳定性。对于设计效应矩阵估计值,我们构造了一致(加权的)组群比例估计值向量的二项协方差矩阵估计值。这样的设计效应矩阵估计值,主要用来消除检验和建模过程中的群内相关。第 7 章和第 8 章将广泛讨论它。使用非加权的比例向量估计值的二项协方差矩阵估计值可以得到另一个设计效应矩阵估计值。它纳入了诸如加权过程的所有其他复杂抽样对协方差矩阵估算的影响作用。我们将在章节 9.3 与 9.4 中,用实例来演示这两种方法。应当注意到,调查分析软件可以使用不同的设计效应的定义。这将导致同一数据的不同设计效应估计值。因此,应当小心不要胡乱解释。

更多的文献

在沃尔特(Wolter, 1985)中,可以找到比率和其他非线性估计值的方差估算的深入讨论。除了已经提及的论题外,凯尔顿(Kalton, 1983)与维尔玛等(Verma et al., 1980)提供了补充材料。基什(Kish, 1995)给出了设计效应概念的完整讨论。邵和涂(Shao and Tu,1995)则讨论了折刀与脱靴技术。拉奥与邵(Rao and Shao,1993)以及扬与拉奥(Yung and Rao, 2000)讲述了方差估

算的折刀技术。拉奥(Rao, 1999)评述了复杂抽样情形下方差估算的高级论题。

斯金纳等(Skinner et al. , 1989)讨论了组群比率估计值向量的渐进协方差矩阵的估算。辛格(Singh, 1985)、库马尔与辛格(Kumar and Singh, 1987)、莫雷尔(Morel, 1989)以及雷同能(Lehtonen, 1990)推导了不稳定情形下的平滑估计值。斯科特(Scott, 1986)介绍了,拉奥与斯科特(Rao and Scott, 1992)运用了有效样本规模的方法。布赖尔(Brier, 1980)、威廉姆斯(Williams, 1982)及威尔逊(Wilson, 1989)讨论了使用 β 二项抽样模型来消除额外二项离异。文献中对复杂调查分析中的不等选中概率的加权以及无应答的校正相当关注。其中,利特尔(Little, 1991,1993)、基什(Kish, 1992)、普费弗曼(Pfeffermann, 1993),以及普费弗曼等(Pfeffermann et al. ,1998)作出了重要的贡献。

组群的模型辅助估算

Model-Assisted Estimation for Domains

第6章

在本章中，我们检视总体子群或组群的估算。以行政标准划分的地域，如县或自治市是典型的组群或重要组群。在社会调查中，总体也可以按照人口标准分成诸如性别和年龄组的组群。在商业调查中，企业通常根据其所在的行业分成组群。另外，地域中的元素也可以根据人口标准分配到不同的组群中。在所有这些例子中，组群估算是指诸如总和的目标子群中的总体数量的估算。我们将在基于设计的估算场合讨论组群总和的估算，这也是本书主要的方法。在实际中，基于设计的估算主要用于样本规模较大的组群。对于小组群（组群内的样本规模较小的），则通常使用小规模估算的模型。在章节6.1，我们给出组群估算的框架和基本原则。同时，我们也归纳出组群估算程序的操作性步骤。章节6.2中，我们介绍两个重要的概念，在组群估算场合下的估计值类型与模型选择。章节6.3举出并讲解某些估计值与模型。章节6.4用实例来讨论基于蒙特卡洛实验的组群总和估计值的特征。章节6.5给出小结与文献。

6.1 组群估算的框架

我们主要关注描述性调查中组群的总体总和估算。借助辅助信息，我们从基于设计的角度来讨论组群总和的估算。根据桑德尔等（Särndal et al., 1992），这一框架被称为模型辅助。在组群估算的过程中，纳入辅助资料的理由是明显的：与不使用辅助资料的估算相比，使用强有力的辅助资料可以得到更为准确的组群估计值。所以，本章扩展了章节3.3中介绍的模型辅助的估算方法。

在模型辅助的估算中，可以使用不同类型的辅助资料。在章节3.3中，我们使用了总体层次的辅助变量的合计。这里，我们也将在组群的模型辅助估算中，使用元素层次的辅助资料。通过使用元素层次的统计模型，将这些资料

纳入到组群估算的程序当中。如果我们设定如下的假设前提,这些是可行的:(1)可以使用登记资料(例如人口普查登记、商业登记、各种行政登记)作为总体框架与辅助资料来源;(2)登记资料含有独特的标签可以用来在个体层次合并登记资料与样本调查数据(见第 1 章中的图 1.1)。显然,能够得到个体层次合并了的登记资料与调查资料,将极大地提高组群估算程序中的灵活性。桑德尔(Särndal,2001)与雷同能等(Lehtonen et al.,2003)采用了这样的观点。本章的许多资料都来源于此。

小范围估算特有的方法包括多种模型决定性技术,诸如合成(SYN)估算公式、复合估算公式、最佳经验线性预测估算公式(EBLUP)与各种贝叶斯技术,以及在人口与疾病预测中发展出来的技术。J. N. K. 拉奥(J. N. K. Rao,2003)给出了模型决定性的小规模估算详细的讲解,并讨论了组群估算的方法论。其他有谢伯理(Schaible,1996),劳森等(Lawson et al.,1999)以及戈什(Ghosh,2001),他们特别讨论了经验与嵌套的贝叶斯技术。

基本原则

让我们引入组群估算中总体量与样本量的基本符号。有限总体还是用 $U=\{1, 2, \cdots, k, \cdots, N\}$ 来表示。同时,在组群估算中,我们使用一组互斥的总体子群,用 $U_1, \cdots, U_d, \cdots, U_D$ 来表示(注意,在本章我们特别使用下注 d 来表示目标组群)。我们假定总体 U 可以用作抽样框。这意味着 U 是一个计算机化的数据,如人口登记或是企业登记。所以,我们同时假定,对于所有的 $k \in U$ 的元素,框架总体 U 包含其他变量(除去总体元素的"标签"k 以外)的取值(符号"$\in$"表示一个元素属于一个元素集)。这些变量是独特的元素标签(ID),组群集的标识,层级标识,以及辅助 z-变量。

用 y 来表示目标变量,Y_k 来表示元素 k 的总体取值。目标参数是组群总和,$T_d = \sum_{k \in U_d} Y_k, d = 1, \cdots, D$,其中的求和是针对所有属于组群 U_d 的总体元素 k(为了简便,我们在本章中使用这样的符号)。在构建准确的组群估算值时,辅助变量非常重要;在样本规模变小时,更为重要。令 $\mathbf{z}_k = (z_{1k}, \cdots, z_{jk}, \cdots, z_{Jk})'$为维度 $J \geqslant 1$ 的辅助变量向量。对于每一个 $k \in U$ 的元素,$\mathbf{z}_k$ 的取值假定已知。在个人的调查中,$\mathbf{z}_k$ 可以是个人 k 的一致数据,如,年龄、性别、征税收入以及其他连续或是定类变量取值。在商务调查中,$\mathbf{z}_k$ 可以表示公司 k 的离职人数或者是员工总数。需要强调的是,我们假定 z 的辅助资料应当在微观个体层次,亦即框架登记中的每一个总体元素都被赋值。这是为了灵活性。这样的话,如果需要,数据可以加总到总体的更高层次,如组群或是层级。事实上,知道每一个目标组群的辅助变量 z_j 的总体总和 $T_{dz1}, \cdots, T_{dzj}$就足够了。在模型拟合阶段,我们通常假定常数 1 为向量 $\mathbf{z}_k$ 的第一个元素。

对于每一个总体元素独特的组群标识,我们定义 $\boldsymbol{\delta}_k = (\delta_{1k}, \cdots, \delta_{dk}, \cdots,$

δ_{Dk})′为单位 k 的组群标识向量。其中,对于所有的 $k \in U_d$ 的元素,$\delta_{dk}=1$,而对于所有的 $k \notin U_d$ 的元素,$\delta_{dk}=0$,$d=1, \cdots, D$。总体元素 k 的另一个层级标识向量 $\boldsymbol{\tau}_k$ 也有相似的定义:对于所有的 $k \in U_h$,$\tau_{hk}=1$,而对于所有的 $k \notin U_h$,$\tau_{hk}=0$,$h=1, \cdots, H$。U_h 是指层级 h,H 则为层级的个数。这样,在总体框架中假定了 D 个组群标识变量,H 个层级标识变量。

使用抽样设计 $p(s)$ 从 U 中抽取规模为 n 的概率样本 s,使得每一单位 k 的选中概率为 π_k。相应的抽样权重为 $w_k=1/\pi_k$。对于 $k \in U$ 中的元素,得到回应变量 y 的测量值 y_k。我们假定样本 s 中含有独特的元素标识,以便于将这些数据与登记框架 U 相合并。

组群样本为 $s_d=U_d \cap s$,$d=1, \cdots, D$。如果在抽样设计中,组群样本规模 n_{sd}没有确定,则定义组群为非设计的。这发生在目标组群结构没有纳入抽样设计的情形中。这样的话,组群样本规模是增加组群估计值方差的随机量。另外,如果总体中某一组群的规模较小,其样本元素数目可能很小(甚至为0)。另一方面,对于设计的组群,根据分层,组群样本规模为事先确定。在实际中,通常使用有着恰当配额方案的分层抽样。

表 6.1 给出了一个 n 元素的分层样本的组群结构。在表格形式下,非设计的组群结构横跨层级。这在实际中经常遇到。在其他类型的结构中,层级与组群可以互为嵌套。例如,一个层级含有几个非设计组群(就像更大区域中的几个次级区域),或者几个层级本身构成组群。后一种情形即是一种设计的组群。辛格等(Singh et al. ,1994)详细地讲解了组群估算中的设计组群的方法。在加拿大劳动力调查中,他们给出了折中抽样配额方案,以满足省级与省级内两个层次上的信度要求。但是,由于实际上的原因,通常无法将目标组群定义为层级。

表 6.1 n 个元素的分层样本中设计与未设计的组群结构

未设计的组群	层级(设计的组群)						
	1	2	…	h	…	H	总和
1	$n_{s_{11}}$	$n_{s_{12}}$	…	$n_{s_{1h}}$	…	$n_{s_{1H}}$	n_{s_1}
2	$n_{s_{21}}$	$n_{s_{22}}$	…	$n_{s_{2h}}$	…	$n_{s_{2H}}$	n_{s_2}
⋮	⋮	⋮		⋮		⋮	⋮
d	$n_{s_{d1}}$	$n_{s_{d2}}$	…	$n_{s_{dh}}$	…	$n_{s_{dH}}$	n_{s_d}
⋮	⋮	⋮		⋮		⋮	⋮
D	$n_{s_{D1}}$	$n_{s_{D2}}$	…	$n_{s_{Dh}}$	…	$n_{s_{DH}}$	n_{s_D}
总和	n_1	n_2	…	n_h	…	n_H	n

样本规模 n_{s_d},$d=1,\cdots,D$,因为未设计组群并不事先确定,因此其为随机变量。

层级样本规模 n_h,$h=1,\cdots,H$ 在抽样设计中是确定的,因而层级是设计的组群。

方格样本规模 $n_{s_{dh}}$在两种情形下均是随机变量。

在组群估算中，应当尽可能使用设计组群方法。在这一过程中，定义最重要的目标组群为层级，并在抽样设计中使用恰当的诸如指数或是班基尔配额方案（见下一个例子）。如果运用非设计组群方法，使用大规模样本来避免可能的组群规模过小，也是有益的。在估算阶段，通常使用精心选择的模型与组群总和的估算公式，将强有力的辅助数据纳入估算程序中（见范例6.2与章节6.4）。

范例6.1

抽样设计对组群估算的影响作用：设计与非设计组群结构。在使用非设计组群结构过程中，当整体样本规模较小时，由于可能抽中总体规模较小的小组群样本，因而遇到并不准确的估算。例如，如果使用无放回式简单随机抽样抽取样本，则某一组群的样本规模为 $E(n_{sd}) = n \times (N_d / N)$，与分层抽样中的按比例配额相同。替代方案是基于设计组群的结构，将组群定义为层级。这样，可以使用更恰当的配额方案。在这个例子中，配额方案是基于指数配额（见章节3.1）。在指数或是班基尔配额中，样本在组群中的配额是基于组群中的回应变量 y 的离异系数信息与可能已知的辅助变量 z 的组群总和 T_{dz} 来决定的。我们使用假设情形下的指数配额的简化形式。其中，回应变量 y 的离异系数 $C.V_{dy} = S_{dy}/\bar{Y}_d$ 对于所有组群已知，S_{dy} 与 $\bar{Y}_d$ 分别是组群 d 里的 y 的总体标准差与总体均值。

在指数配额中，组群的样本规模为，

$$n_{d,pow} = n \times \frac{T_{dz}^{a} \times C.V_{dy}}{\sum_{d=1}^{D} T_{dz}^{a} \times C.V_{dy}}$$

其中，系数 a 是指期望的指数（通常取值为0，0.5或者1）。这里，我们为了简便，选择 $a=0$。所以，只使用了离异系数的信息。

我们从职业健康保健调查（OHC）数据（$N=7\ 841$ 人）中选取一个SRSWOR样本（$n=392$ 人），并估算构建的 $D=30$ 个组群中的慢性病患者的总数。通过这些，我们来讲解方法。在总体中，组群的规模从最少81人到最多517人。表6.2给出了使用按比例配额方案（与非设计组群结构相应）与使用指数配额方案（与设计组群结构相应）的样本。慢性病患者的组群总数用霍维茨-汤普森（HT）公式 $\hat{t}_{dHT} = \sum_{k \in sd} w_k y_k$ 来估算。估算公式的稳定性由组群总和估计值的总体离异系数 $C.V(\hat{t}_{dHT}) = S.E(\hat{t}_{dHT})/T_d$ 来表示。

表 6.2　OHC 调查数据中元素 $n=392$，组群 $D=30$ 的样本配额方案

组群		组群样本规模		组群合计的 HT 估算值的离异系数(%)	
		未设计组群结构 SRSWOR 期望值	设计的组群结构分层 SRSWOR 结果(指数配额)	未设计的组群结构 SRSWOR	设计的组群结构分层 SRSWOR (指数配额)
d	N_d	$E(n_{s_d})$	n_d	$C.V(\hat{t}_{dHT})$	$C.V(\hat{t}_{dHT})$
10	81	4	11	84.10	38.88
20	101	5	12	78.41	40.54
18	129	6	13	72.69	42.38
3	133	7	15	81.04	45.63
8	141	7	16	81.03	46.54
30	146	7	15	74.80	45.03
21	153	8	12	62.87	41.15
23	156	8	11	57.65	39.05
16	165	8	13	64.94	43.19
1	181	9	17	75.90	48.78
11	187	9	14	63.52	44.52
6	188	9	13	60.37	43.22
28	194	10	10	50.52	38.69
24	200	10	13	58.68	43.39
22	242	12	10	44.27	38.30
15	252	13	14	55.68	45.50
7	292	15	17	60.34	50.06
4	295	15	15	53.92	47.04
13	305	15	13	46.00	43.04
12	311	16	12	44.50	42.38
5	323	16	16	53.50	48.23
25	339	17	11	40.57	41.03
2	352	18	14	46.80	45.74
26	364	18	11	38.87	40.88
29	365	18	11	38.25	40.45
9	366	18	14	45.99	45.85
17	426	21	12	36.67	41.62
14	447	22	13	37.95	43.37
19	490	24	11	33.60	41.22
27	517	26	10	30.68	39.34
总和	7 841	392	392		

数据来自：SRSWOR 设计计算组群样本期望值，指数配额($a=0$)分层 SRSWOR 设计计算组群样本规模的结果，以及相应的霍维茨-汤普森估计值的离异系数(%)。

结果显示,SRSWOR 抽样在期望组群样本规模上产生了较大的离异:平均组群规模为 13,最小为 4,最大为 26。另一方面,指数配额在很大程度上抹平了组群样本规模的离异:最小为 10,最大为 17。在 SRSWOR 情形下,离异的百分比系数变动很大。比如,最小与最大的离异系数之间的差异超过了 60%点。在指数配额中,这一差异减少为 12%点。因此,指数配额因减小较大的系数而抹平了离异系数。但是,估算的组群总和的离异系数比较大,这主要是因为整体上的样本规模较小。

离异系数的改善可以用图形显示出来。在图 6.1 中,根据总体组群规模大小变化画出离异系数。在 SRSWOR 情形下得到的 HT 估计值的曲线明显随着组群规模增加而减小,而指数配额的估计值的曲线则明显稳定。

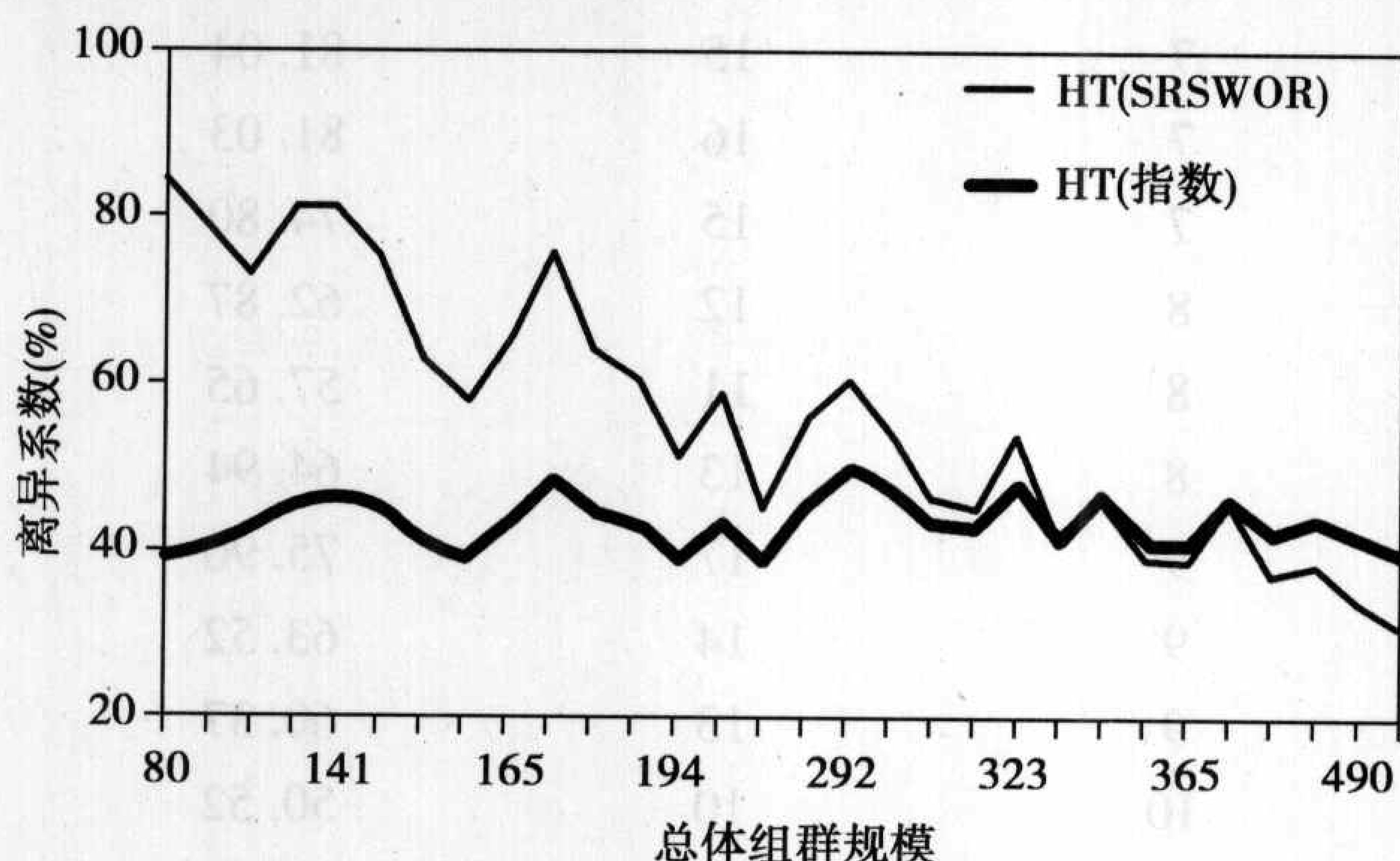


图 6.1 SRSWOR 抽样(对应于未设计组群结构)与指数配额($a=0$)分层 SRSWOR 抽样(对应于设计的组群结构)下的组群总和的霍维茨-汤普森估算值的离异系数(%)

为了进一步指明组群估算的条件,我们有如下技术假设。我们假定,在收集选中样本的资料并准备最终样本数据(用 $s(y)$ 来表示)后,总体框架 U 与样本测量值 $s(y)$ 可以通过两个数据都有的独有的元素标签 ID 合并起来。完成这一步骤后,我们得到一个扩充的框架登记数据,其中,总体中所有元素含有 z 的辅助数据与层级和组群标识变量,而样本中的元素另外还有 y 的测量值。

至此,我们完成了组群估算的技术准备。资料 6.1 以更为普通的方式给出了组群估算的操作步骤。

资料 6.1 组群估算程序的操作步骤

第一步:构建总体框 对于 U 中的所有元素 k,构造含有独特标签的 N 个元素的框架总体 $U=\{1, 2, \cdots, k, \cdots, N\}$,组群标识向量 $\boldsymbol{\delta}_k$,层级标识向量 $\boldsymbol{\tau}_k$,应用抽样设计 $p(s)$ 抽取 n 个元素样本的选中概率 π_k,以及 z 的辅助资料向量 z_k。

第二步:抽样与测量　对于所有 $k \in s$ 的元素,使用 $p(s)$ 抽取样本,并测量回应变量 y 的取值,构建包括元素标签 ID、观测值 y_k 以及抽样权重 $w_k = 1/\pi_k$ 的样本数据 $s(y)$。

第三步:重访框架总体　使用元素标签 ID,将框架总体 U 与样本数据 $s(y)$ 合并起来,以构造一个新的组合数据。

第四步:模型选择与模型拟合　选择模型,指明模型参数与效应,应用样本数据来拟合、检验及诊断模型。根据拟合模型,对于所有 $k \in U$ 的元素,计算拟合值 $\hat{y}_k$;对于所有 $k \in s(y)$ 的元素,计算残差 $\hat{e}_k = y_k - \hat{y}_k$。

第五步:选择组群总和估算公式与组群估算　为确定的组群总和估算公式提供拟合值、残差及权重。基本上,"模型决定"的组群总和估算公式使用 $k \in U$ 的拟合值 $\hat{y}_k$;"模型辅助"的组群总和估算公式使用 $k \in U$ 的拟合值 $\hat{y}_k$,以及 $k \in s$ 的残差 $\hat{e}_k$ 与权重 w_k。

第六步:方差估算与诊断　选择恰当的方差估算公式。计算标准误估计值与离异系数。

表 6.3　资料 6.1 中组群估算过程第 1,3,4 步的执行情况(假设的情形)

第 1 步:构建总体框 U					第 3 步:将总体框 U 与样本数据 $s(y)$ 合在一起			第 4 步:计算拟合的 y 值与残差	
元素 *ID*	组群 *ID* 向量 δ'_k	层级 *ID* 向量 τ'_k	选中概率 π_k	辅助 z 向量 z'_k	抽样权重 w_k	样本元素标识 I_k	研究变量 y_k	拟合值 $\hat{y}_k$	残差 $\hat{e}_k$
1	δ'_1	τ'_1	π_1	z'_1	0	0	…	$\hat{y}_1$	…
2	δ'_2	τ'_2	π_2	z'_2	0	0	…	$\hat{y}_2$	…
3	δ'_3	τ'_3	π_3	z'_3	w_3	1	y_3	$\hat{y}_3$	$\hat{e}_3$
4	δ'_4	τ'_4	π_4	z'_4	0	0	…	$\hat{y}_4$	…
5	δ'_5	τ'_5	π_5	z'_5	w_5	1	y_5	$\hat{y}_5$	$\hat{e}_5$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	δ'_k	τ'_k	π_k	z'_k	w_k	1	y_k	$\hat{y}_k$	$\hat{e}_k$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	δ'_N	τ'_N	π_N	z'_N	0	0	…	$\hat{y}_N$	…

…:未被抽中的元素。

在表6.3中,我们总结了一个假设的情形,其中有了资料6.1第1到第4步中组群估算程序在一个总体框数据的执行情况。对于包括抽取或是未抽取的所有元素而言,由于辅助变量 z 的向量 $\mathbf{z}_k = (z_{1k}, \cdots, z_{jk}, \cdots, z_{Jk})'$ 为已知,辅助变量 z 的总体总和向量 $\mathbf{T}_{z_j} = \sum_{k \in U} z_{jk} (j = 1, \cdots, J)$ 的 $\mathbf{T}_z = (T_{z_1}, \cdots, T_{z_J})'$ 也是已知。同时,因为对于所有 $k \in U$ 的组群标识已知,对于每一个变量 z 的组群总和

$T_{d_j} = \sum_{k \in U_d} z_{jk} (d = 1, \cdots, D$ 及 $j = 1, \cdots, J)$ 可以计算出来。整个总体数据生成了一个样本标识变量 I。当 $k \in s$ 时，$I_k = 1$，其余情形为 0。显然，总体中的标识变量的和为 n。在模型拟合阶段，所有 $k \in U$ 的 N 个元素均计算拟合值 $\hat{y}_k$。另一方面，仅计算 $k \in s$ 的残差 $\hat{e}_k = y_k - \hat{y}_k$。应当强调，给定模型计算出的拟合值 $\{\hat{y}_k;\ k \in U\}$ 与另一个模型计算所得不同。在下个小节进一步讲解组群总和的模型与估算中，这一点将更加清楚。

6.2 估算类型与模型选择

组群估算过程中的重要步骤包括：选择总和的估算公式类型，选择使用的辅助变量，将辅助资料纳入估算程序中的模型建立、模型拟合，以及选中组群总和估算公式的方差估计值的推导（见资料 6.1）。在本章节中，我们用更为技术化的方式来讨论这些。

为了为构造目标组群的总体总和的估算公式打下基础，我们首先讨论两个概念：估算公式类型和模型选择。

估算公式类型

估算公式类型是指选中组群总和的估算公式的明确结构。本章讨论的主要有两种估算公式类型，通用回归（GREG）估算公式与合成（SYN）估算公式。这两个估算公式的概念上的主要差异是，GREG 估算公式用模型作为辅助工具，而 SYN 估算公式则完全依赖于所用的模型。因此，GREG 估算公式是模型辅助性的，而 SYN 是模型决定性的。模型的不同角色的主要后果是，组群总和的 GREG 估算公式被构造为设计无偏差（或者接近于此），而不管模型“正确”与否。这是 GREG 估算公式的优势。但是，当组群的样本规模较小时，GREG 估算公式可以很不稳定。另一方面，SYN 估算公式的偏差极大地依赖于模型的正确构建。当模型构建偏误严重，SYN 估算公式可以有相当的设计偏差。但是，当模型构建正确合理时，SYN 估算公式的偏差可以很小。

在例如由全国性统计机构实施的典型的大规模调查中，某些目标组群的规模足够大，辅助信息足够有力，因而 GREG 类型的估算公式足够精确。但对于较小的组群规模，GREG 估算值的方差可能非常大，而同样例子中的 SYN 估算公式的方差则很小。由于 SYN 估算值在小组群中的精确性，使得它更多地用于小区域估算中（回忆一下，“小区域”是指给定组群的样本规模或是“区域”很小，甚至接近于 0）。

小结一下估算公式类型的主要理论特征，GREG 估算公式构造成设计无偏差，SYN 估算公式则不是如此。当组群样本规模较小时，GREG 估算值的方

差对于小组群而言,可能较大,导致较差的精度。SYN 估算公式通常是有设计偏差的;当样本规模增加时,其偏差并不趋近于 0;其方差通常小于 GREG 估算值;并且它是特别针对小组群的。当偏差较大时,即使方差较小,由平均方差 MSE 表示的 SYN 估算值的精度可能较差。

模型选择

模型选择的概念是指,明确研究变量 y 与辅助预测变量 $z_1,\cdots,z_J$ 之间在构造的模型的结构中所反映出来的关系。模型选择有两个方面,模型的数学形式以及模型中参数与效应的明确化。例如,研究变量是一个连续变量时,线性模型的形式通常是恰当的。对于二分变量或是多值变量,可以选择非线性模型,诸如二分或多值对数模型。又如,对于二分研究变量,对数回归模型可能比线性模型更佳。这是因为前者的拟合值正好落入单位区间内,而这对于线性模型则并不往往如此。

模型选择的第二个方面就是模型中参数与效应的指明。其中的一些是在加总的总体层次,一些是在组群层次(组群层次的参数),另一些则在其间的层次。我们将区分固定模型与混合模型。固定模型包含总体层级、组群层级或是其间层级的固定效应,在混合模型中,除了固定效应外,还有组群间的随机效应。通过使用混合类型的模型,我们可以引入随机效应来区分组群间的差异。

小结一下,选中的模型指明了研究变量 y 与预测变量 $z_1,\cdots,z_J$ 之间的理论假设关系,并假定其可能很复杂的残差结构。通常固定模型就可以了,但混合模型在建模中提供了更多的灵活性。对于每一个选定的模型,根据不同的建模原则,我们能够推导出一个 GREG 与 SYN 估算值。但是,固定模型更常见于模型辅助估算公式中,而混合模型更常见于模型决定性估算公式中。

结合组群总和估算公式中的两个方面,估算公式类型与模型选择,我们得到一个两维的估算情形。为了说明这一情形,我们在表 6.4 中介绍一些估算值。表中有 6 种模型决定性 SYN 类型估算值和 6 种基于设计的 GREG 类型估算值。总体模型(P-模型;第 1,2 行)中唯一的参数为定义在总体层次的固定效应;它不含组群层次的参数。组群模型(D-模型)至少含有定义在组群层次的一些参数或效应。这些是第 3,4 行的固定效应,以及第 5,6 行的随机效应。“线性”与“对数”是指数学形式。范例 6.2 和章节 6.4 中,我们将更详细地讨论这些估算公式。

表 6.4　分模型选择与估算值类型的组群总和估算公式列表

模型选择			估算值类型	
模型效应	加总层次	函数形式	模型决定性	基于设计的模型辅助
固定效应模型	总体模型	线性	SYN-P	GREG-P
		对数	LSYN-P	LGREG-P
	组群模型	线性	SYN-D	GREG-D
		对数	LSYN-D	LGREG-D
混合效应模型（固定与随机效应）	组群模型	线性	MSYN-D	MGREG-D
		对数	MLSYN-D	MLGREG-D

6.3　估算值的构造与模型设定

组群总和估算值的构造

组群总和估算值通过以下 3 个步骤构造（根据资料 6.1 中的 3,4,5 步）：

1. 根据样本数据 $s(y)=\{(y_k,\ \mathbf{z}_k);\ k\in s\}$ 估计设定的模型参数。
2. 使用模型参数的估计值与总体向量 $\mathbf{z}_k$，对于每一个属于与不属于样本的元素，计算拟合值 $\hat{y}_k$。
3. 为了得到组群 d 的总和 T_d 的估计值 $\hat{t}_d$，将拟合值 $\{\hat{y}_k;\ k\in U\}$ 与样本观测值 $\{y_k;\ k\in s\}$ 纳入相应的 GREG 与 SYN 估算公式中。

我们将结合线性模型的例子说明组群估算的程序。考虑一个固定效应线性模型，$y_k=\mathbf{z}_k'\boldsymbol{\beta}+\varepsilon_k$，其中 $\boldsymbol{\beta}$ 为将要估算的未知参数，ε_k 为残差项。拟合模型得到 $\hat{\boldsymbol{\beta}}$。对于所有 $k\in U$ 的元素，拟合值为 $\hat{y}_k=\mathbf{z}_k'\hat{\boldsymbol{\beta}}$。相似的，对于一个在固定效应之外含有组群层次随机效应的线性混合模型，模型为 $y_k=\mathbf{z}_k'(\boldsymbol{\beta}+\mathbf{u}_d)+\varepsilon_k$，其中的 $\mathbf{u}_d$ 是定义在组群层次的随机效应向量。使用参数估计值，所有 $k\in U$ 的元素的拟合值为 $\hat{y}_k=\mathbf{z}_k'(\hat{\boldsymbol{\beta}}+\hat{\mathbf{u}}_d)$。用更加概括的语言，估算组群总和的 GREG 与 SYN 模型构造是通用性线性混合模型的特例，例如为一个混合线性模型与一个对数模型（McCulloch and Searle，2001；Dempster et al.，1981）。

模型的构造不同，则拟合值 $\{\hat{y}_k;\ k\in U\}$ 各不相同。给定一个模型，两个基本估算值类型的组群总和 $T_d=\sum_{k\in U_d}y_k$ 的估算公式有以下结构：

合成估算公式：

$$\hat{t}_{dSYN} = \sum_{k \in U_d} \hat{y}_k \tag{6.1}$$

通用回归估算公式：

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} w_k(y_k - \hat{y}_k) \tag{6.2}$$

其中，$w_k = 1/\pi_k$，$s_d = s \cap U_d$是整个样本 s 中落入 U_d 的部分，$d = 1, \cdots, D$。

注意，$\hat{t}_{dSYN}$ 使用估算模型的拟合值，因而依赖于模型的"真实性"，所以可能有偏差。另一方面，$\hat{t}_{dGREG}$ 有一项用来防备可能的模型设定错误。同时注意，在某一组群没有样本元素的情形下，该组群的$\hat{t}_{dGREG}$ 简化为$\hat{t}_{dSYN}$。在评估更为复杂的估算值时，通常将霍维茨-汤普森估算值 $\hat{t}_{dHT} = \sum_{k \in s_d} w_k y_k$ 当成参照对象。

模型设定

让我们首先讨论固定效应线性模型。令 $\mathbf{z}_k = (1, z_{1k}, \cdots, z_{jk}, \cdots, z_{Jk})'$为一个$(J+1)$维的、含有预测变量 z_j 取值的向量，$J \geqslant 1$，$j = 1, \cdots, J$。这一向量用来生成估算公式 6.1 与式 6.2 预测值$\hat{y}_k$，$k \in U$。

1. 固定效应 P-模型。SYN-P 与 GREG-P 估算值是建立在以下模型之上：

$$y_k = \beta_0 + \beta_1 z_{1k} + \cdots + \beta_J z_{Jk} + \varepsilon_k = \mathbf{z}'_k \boldsymbol{\beta} + \varepsilon_k \tag{6.3}$$

对于 $k \in U$，$\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_J)'$是在整个总体上定义的固定效应向量。根据这样特征，我们把式 6.3 称为固定效应 P-模型。当观测到整体的 y-数据，我们可以计算 $\boldsymbol{\beta}$ 的通用最小二乘法（GLS）估算值：

$$\mathbf{B} = \left(\sum_{k \in U} \mathbf{z}_k \mathbf{z}'_k / c_k \right)^{-1} \sum_{k \in U} \mathbf{z}_k y_k / c_k, \tag{6.4}$$

其中，c_k 是设定的正的权重。在不显著失去概括性的情形下，对于 $k \in U$，我们设定 $c_k = \boldsymbol{\lambda}' \mathbf{z}_k$，其中$(J+1)$维的 $\boldsymbol{\lambda}$ 与 k 无关。进一步简化，对于所有的 k，我们可以设定 $c_k = 1$，式 6.4 简化为普通最小二乘法（OLS）估算值。在实际中，我们使用观测到的样本数据来计算式 6.4 的加权最小二乘法（WLS）估计值，得到

$$\hat{\mathbf{b}} = \left(\sum_{k \in s} w_k \mathbf{z}_k \mathbf{z}'_k \right)^{-1} \sum_{k \in s} w_k \mathbf{z}_k y_k, \tag{6.5}$$

其中，$w_k = 1/\pi_k$ 是单位 k 的抽样权重。预测值为：

$$\hat{y}_k = \mathbf{z}'_k \hat{\mathbf{b}}, k \in U。 \tag{6.6}$$

将预测值$\hat{y}_k$ 纳入式 6.1 与式 6.2，我们得到相应的 SYN-P 与 GREG-P 估算值。注意，对给定的 d 组群使用 P-模型，其他组群的 y 值也影响纳入 SYN-P 与 GREG-P 估算公式中的该预测值。由于这一原因，运用固定效应 P-模型的估算值$\hat{t}_{dSYN\text{-}P}$与$\hat{t}_{dGREG\text{-}P}$被称为间接估算值。

2. 固定效应 D-模型。SYN-D 与 GREG-D 估算值同样是建立在预测变量

$\mathbf{z}_k$ 之上的。但是,模型的设定不同,允许了各个组群有一个固定效应 $\boldsymbol{\beta}_d$。因此,对于 $k \in U_d, d=1, \cdots, D$,有

$$y_k = \mathbf{z}'_k \boldsymbol{\beta}_d + \varepsilon_k \tag{6.7}$$

或是,相同的,对于 $k \in U$,

$$y_k = \sum_{d=1}^{D} \delta_{dk} \mathbf{z}'_k \boldsymbol{\beta}_d + \varepsilon_k \tag{6.8}$$

其中,δ_{dk}是单位 k 的组群标识变量。定义为,$k \in U_d, \delta_{dk}=1$,当 $k \notin U_d, \delta_{dk}=0, d=1, \cdots, D$。式 6.7 被称为固定效应 D-模型。当式 6.7 拟合整个次级总体 U_d 的数据时,β_d 的 GLS 估算值为,

$$\mathbf{B}_d = \left(\sum_{k \in U_d} \mathbf{z}_k \mathbf{z}'_k / c_k\right)^{-1} \sum_{k \in U_d} \mathbf{z}_k y_k / c_k, d = 1, \cdots, D. \tag{6.9}$$

在实际中,拟合必须依据观测到的组群 d 中样本数据。对所有的 k,设定 $c_k=1$,可以使用以下 WLS 估算值:

$$\hat{\mathbf{b}}_d = \left(\sum_{k \in s_d} w_k \mathbf{z}_k \mathbf{z}'_k\right)^{-1} \sum_{k \in s_d} w_k \mathbf{z}_k y_k, d = 1, \cdots, D. \tag{6.10}$$

其对于 $k \in U_d, d=1, \cdots, D$,预测值为,

$$\hat{y}_k = \mathbf{z}'_k \hat{\mathbf{b}}_d \tag{6.11}$$

将式 6.11 的预测值 $\hat{y}_k$ 纳入式 6.1 与式 6.2 中,我们得到相应的 SYN-D 与 GREG-D 估算值。对给定的组群 d,在模型拟合以及计算纳入该组群 SYN-D 与 GREG-D 估算值的预测值的过程中,只有该组群的 y- 值。因此,固定效应 D- 模型估算值 $\hat{t}_{dSYN\text{-}D}$ 与 $\hat{t}_{dGREG\text{-}D}$ 被称为直接估算值。注意,由于设定 $c_k = \boldsymbol{\lambda}'\mathbf{z}_k = 1$,我们有 $\sum_{k \in s_d} w_k (y_k - \hat{y}_k) = 0$。相应的,SYN-D 与 GREG-D 相等,即,对每一个样本 s,当使用固定效应 D-模型时,$\hat{t}_{dSYN\text{-}P} = \hat{t}_{dGREG\text{-}P}$。

3. 混合 D-模型。MSYN-D 与 MGREG-D 估算值是建立在叫做混合线性 D-模型的两级线性模型之上的,其中包括固定效应以及表示组群差异的随机效应,对于 $k \in U_d, d=1, \cdots, D$,

$$\begin{aligned} y_k &= \beta_0 + u_{0d} + (\beta_1 + u_{1d}) z_{1k} + \cdots + (\beta_J + u_{Jd}) z_{Jk} + \varepsilon_k \\ &= \mathbf{z}'_k (\boldsymbol{\beta} + \mathbf{u}_d) + \varepsilon_k \end{aligned} \tag{6.12}$$

每一个系数是固定部分与单个组群的随机部分之和:$\beta_0 + u_{0d}$为截距,$\beta_j + u_{jd}$为斜率,$j=1, \cdots, J$。$\mathbf{u}_d = (u_{0d}, u_{1d}, \cdots, u_{Jd})'$表示对模型中固定效应系数的偏离,

$$y_k = \beta_0 + \beta_1 z_{1k} + \cdots + \beta_J z_{Jk} + \varepsilon_k = \mathbf{z}'_k \boldsymbol{\beta} + \varepsilon_k, \tag{6.13}$$

这与式 6.3 是一致的。更普遍的,我们仅设定式 6.12 中的一些系数为随机,因此,对于某些 j,每一个 d,$u_{jd}=0$。经常在实际中使用的式 6.12 中的简单而又特别的例子是,只有截距项 u_{0d}为随机项,得到 $y_k = \beta_0 +$

$u_{0d}+\beta_1 z_{1k}+\cdots+\beta_J z_{Jk}+\varepsilon_k$。将拟合的 y-值，

$$\hat{y}_k = \mathbf{z}'_k(\hat{\boldsymbol{\beta}}+\hat{\mathbf{u}}_d) \tag{6.14}$$

带入式 6.1，得到两层次的 MSYN-D 估算值。将拟合值式 6.14 带入式 6.2，得到由雷同能与韦建能（Lehtonen and Veijanen，1999）引入的两层次 MGREG-D 估算值。使用极大似然（ML）或是限制性极大似然（REML）来估算方差部分，以及给定方差估计值用 GLS 来估算固定效应，可以拟合一个两层次的 D-模型式 6.12。更详细的内容，见戈尔茨坦（Goldstein，2002）与麦卡洛克与瑟尔（Mcculloch and Searle，2001）的书。在估算混合 D-模型时，通常假定随机效应符合联合正态分布。但是注意，这一正态分布假设在获取近似无偏 MGREG-D 估算值中，并不是必需的。

在估算组群总和式 6.1 与式 6.2 的设计方差中，还有其他方案。对于设计组群，其组群样本规模 n_d 在分层抽样设计中是固定的，当每一层级中样本抽取采用 SRSWOR 时，章节 3.3 中的方差估计值的回归估算方法可以用于各个组群。在这种情形下，对于所有的 $k \in U_d$，从组群 d 的总体 N_d 中抽取 n_d 个样本，权重为 $w_k = N_d/n_d$。例如，式 6.2 中的 GREG 估算值，其近似方差估计值为，

$$\hat{v}_{srs}(\hat{t}_{dGREG}) = N_d^2\left(1-\frac{n_d}{N_d}\right)\left(\frac{1}{n_d}\right)\sum_{k\in s_d}\frac{(\hat{e}_k-\bar{\hat{e}}_d)^2}{n_d-1}, \tag{6.15}$$

其中，残差 $\hat{e}_k = y_k-\hat{y}_k, k\in s_d$，而 $\bar{\hat{e}}_d=\sum_{k\in s_d}\hat{e}_k/n_d$ 是组群 d 的残差平均值，$d=1,\cdots,D$。很明显，在权重为常数的 SRSWOR 情形下，各个组群内直接估算值的残差和为 0。但在其他设计中，间接估算值的残差和可能不等于 0。

在非设计组群的情形下，应当考虑由随机组群样本规模 n_{sd} 引起的额外变动。让我们考虑从总体 N 个元素中以 SRSWOR 形式抽取 n 个元素的情形。对于所有 k，抽样比率为 n/N，权重为 $w_k=N/n$。令 $y_{dk}=\delta_{dk}y_k, \hat{e}_{dk}=y_{dk}-\hat{y}_k$，$d=1,\cdots,D$，其中的组群标识变量设定为，当 $k\in U_d, \delta_{dk}=1$，其余为 0，我们可以得到一个近似的方差估计值，

$$\hat{v}_{srs}(\hat{t}_{dGREG}) = N^2\left(1-\frac{n}{N}\right)\left(\frac{1}{n}\right)\sum_{k\in s}\frac{(\hat{e}_{dk}-\bar{\hat{e}}_d)^2}{n-1}。 \tag{6.16}$$

注意，组群 d 之外的元素也对方差估计值有影响。因为，对于 $k\notin U_d$ 与 $k\in s, \hat{e}_{dk}=-\hat{y}_k$。另一个方差估计值为，

$$\hat{v}_{srs}(\hat{t}_{dGREG}) = N^2\left(1-\frac{n}{N}\right)\left(\frac{1}{n}\right)p_d\sum_{k\in s_d}\frac{(\hat{e}_k-\bar{\hat{e}}_d)^2}{n_d-1}\left(1+\frac{q_d}{\mathrm{c.v}_{d\hat{e}}^2}\right), \tag{6.17}$$

$d=1,\cdots,D$，其中，$p_d=n_d/n, q_d=1-p_d$，而 $\mathrm{c.v}_{d\hat{e}}=\hat{s}_{d\hat{e}}/\bar{\hat{e}}_d$ 是组群 d 中的残差的样本离异系数，而 $\hat{s}_{d\hat{e}}$ 为组群 d 中残差的样本标准差。估算值式 6.17 与贝努利抽样中常用的方差估算值相对应（见范例 2.2）。

让我们进一步讨论,在比率估算与回归估算的情形下,总和估计值的模型选择与估算公式构造。章节3.3讨论了总体总和 T 的估算。在那里,假定已知的总体层次的辅助信息为辅助变量 z 的总和 T_z,辅助固定效应线性回归模型的形式为式6.3给出的 $y_k=\beta_0+\beta_1 z_k+\varepsilon_k, k\in U$。章节3.3给出的总体总和的比率估算值为 $\hat{t}_{rat}=T_z\times\hat{t}/\hat{t}_z$,回归估算值为 $\hat{t}_{reg}=\hat{t}+\hat{b}_1(T_z-\hat{t}_z)$。其中,$\hat{t}$ 与 $\hat{t}_z$ 为 SRSWOR 情形下 T 与 T_z 的估算值,$\hat{b}_1$ 是有限总体回归系数 B_1 的基于样本的 OLS 估计值。针对组群总和 T_d 的估算,可以使用这些比率与回归估算公式。但是,也可以引入更加复杂的模型类别,就如上面讨论过的式6.3、式6.7与式6.12。

考虑到要估算一个连续变量 y 在若干组群 U_d 中的总和 $T_d, d=1, \cdots, D$。假定一个辅助变量 z,可以给出以下辅助模型。

1. y_k 的固定效应 P-模型,$k\in U$:

(1a) $y_k=\beta_0+\varepsilon_k$,同截距模型;

(1b) $y_k=\beta_1 z_k+\varepsilon_k$,同斜率模型;

(1c) $y_k=\beta_0+\beta_1 z_k+\varepsilon_k$,同截距斜率模型。

2. y_k 的固定效应 D-模型,$k\in U_d, d=1, \cdots, D$:

(2a) $y_k=\beta_{0d}+\varepsilon_k$,分组群截距模型;

(2b) $y_k=\beta_{1d} z_k+\varepsilon_k$,分组群斜率模型;

(2c) $y_k=\beta_{0d}+\beta_{1d} z_k+\varepsilon_k$,分组群截距斜率模型。

3. y_k 的混合效应 D-模型,$k\in U_d, d=1, \cdots, D$:

(3a) $y_k=\beta_{0d}+\varepsilon_k=\beta_0+u_{0d}+\varepsilon_k$,分组群随机截距模型;

(3b) $y_k=\beta_{0d}+\beta_1 z_k+\varepsilon_k=\beta_0+u_{0d}+\beta_1 z_k+\varepsilon_k$,分组群随机截距斜率模型。

模型(1b)与(2b)可用于组群的比率估算中,模型(1c)与(2c)则用于回归估算中。很显然,间接 SYN 与 GREG 估算值由模型设定(1)与(3)得到,模型类型(2)则给出直接 SYN 与 GREG 估算值。

例如,使用 P-模型(1b),组群总和 T_d 的估算公式6.1为,

$$\hat{t}_{dSYN\text{-}P}=\sum_{k\in U_d}\hat{y}_k=\sum_{k\in U_d}\hat{b}_1 z_k=T_{dz}\hat{b}_1=T_{dz}\times\hat{t}_{HT}/\hat{t}_{zHT}, d=1,\cdots,D, \tag{6.18}$$

这与总体的比率估算值 $\hat{t}_{rat}$ 很相似。但是,在 $\hat{t}_{dSYN\text{-}P}$ 中使用了组群总和 T_{dz} 而非整体上的总和 T_z。总体斜率 B_1 的估算公式为,

$$\hat{b}_1=\frac{\sum_{k\in s}w_k y_k}{\sum_{k\in s}w_k z_k}=\frac{\hat{t}_{HT}}{\hat{t}_{zHT}},$$

它是两个 HT 估计值——研究变量 y 与辅助变量 z 的总和 $\hat{t}_{HT}$ 与 $\hat{t}_{zHT}$——的比率。这两个估计值是在总体层次计算的。因此,组群总和的估算值是间接的。

在使用整个样本的 y-值时，$\hat{t}_{dSYN\text{-}P}$同时也借力于其他组群。

使用了式 6.18 模型类型的 SYN 估算值可能是有偏的。$\hat{t}_{dSYN\text{-}P}$的偏差近似为，

$$\mathrm{BIAS}(\hat{t}_{dSYN\text{-}P}) = E(\hat{t}_{dSYN\text{-}P}) - T_d \approx - T_{dz}(B_{1d} - B_1),$$

其中，$B_{1d} = \sum_{k\in U_d} y_k / \sum_{k\in U_d} z_k$ 是分组群斜率，$d = 1, \cdots, D$，并且，$B_1 = \sum_{k\in U} y_k / \sum_{k\in U} z_k$ 为总体的斜率。对给定的组群，组群斜率接近于总体斜率时，偏差可忽略。但这些条件不满足时，可能遇到较大的偏差。

相应的组群总和 T_d 的间接 GREG 估算公式 6.2 为，

$$\hat{t}_{dGREG\text{-}P} = \sum_{k\in U_d} \hat{y}_k + \sum_{k\in s_d} w_k(y_k - \hat{y}_k) = \hat{t}_{dSYN\text{-}P} + \sum_{k\in s_d} w_k(y_k - \hat{b}_1 z_k)$$

$$= \hat{t}_{dHT} + \frac{\hat{t}_{HT}}{\hat{t}_{zHT}}(T_{dz} - \hat{t}_{dzHT}) \tag{6.19}$$

近似于总体回归估算公式。但是，其基础模型不同。注意，间接的 GREG 估算公式也尝试“借力于”其他组群。

类型(2b)中的直接 SYN 与 GREG 估算公式，仅使用了给定组群的 y-值。用分组群的$\hat{b}_{1d}$ 来代替$\hat{b}_1$，得到估算公式，

$$\hat{b}_{1d} = \frac{\sum_{k\in s_d} w_k y_k}{\sum_{k\in s_d} w_k z_k} = \frac{\hat{t}_{dHT}}{\hat{t}_{dzHT}}, d = 1, \cdots, D,$$

其中，$\hat{t}_{dHT}$与$\hat{t}_{dzHT}$是组群层次总和 T_d 与 T_{dz}的 HT 估算值。因此，直接估算值 $\hat{t}_{dSYN\text{-}D}$为，

$$\hat{t}_{dSYN\text{-}D} = \sum_{k\in U_d} \hat{y}_k = \sum_{k\in U_d} \hat{b}_{1d} z_k = T_{dz} \hat{b}_{1d}$$

$$= T_{dz} \times \hat{t}_{dHT} / \hat{t}_{dzHT}, d = 1, \cdots, D。 \tag{6.20}$$

对这一模型设定，直接 GREG 对应的$\hat{t}_{dGREG\text{-}D}$与 SYN 估算公式一致。这是因为，GREG 估算公式 6.2 的第二项消失了。

让我们讨论一下估算公式 6.18 与式 6.20 在偏差、精度与准确性上的相对特征。首先，式 6.18 给出的间接估算公式$\hat{t}_{dSYN\text{-}P}$是有偏的，而且，当模型假设不成立时，偏差可能很大。式 6.20 给出的对应的直接估算公式$\hat{t}_{dGREG\text{-}D}$几乎是无偏设计，与模型假设的有效性无关。式 6.18 间接估算值的方差是 n^{-1}阶次，因而在整个样本规模 n 较大时变得很小。另一方面，直接估算值式 6.20 的方差是 n_d^{-1} 阶次，因而在组群 d 的样本规模 n_d 较小时，变得很大。因此，取决于模型的假设以及组群样本规模，这里有一个偏差与精度的交换。利用均方差，$\mathrm{MSE}(\hat{t}_d) = V(\hat{t}_d) + \mathrm{BIAS}^2(\hat{t}_d)$，我们可以得出如下结论。在小组群中，间接估算值式 6.18 比直接估算值式 6.20 更为准确，因为式 6.20 的方差可能较大。

但对于大组群而言(组群样本规模较大),直接估算值更为准确,因为式 6.18 的偏差平方可能占极大的份额。特别在违反模型假设时,这些结论成立(在雷同能等(Lehtonen et al. ,2003)的书中,有这一交换更详细的例子)。

在范例 6.2 中,我们根据从 OHC 调查数据中以 SRSWOR 抽取的单个样本来研究组群总和的某些估算值。在章节 6.4 中,我们更详细地检验不同模型选择情形下的合成与通用回归估算值的相对特征(偏差与准确性)。在那里,将使用蒙特卡洛模拟,即从固定总体中抽取大量独立的 SRSWOR 样本来审视这些方法。

范例 6.2

在 SRSWOR 抽样情形下,基于设计的组群总和估算。我们将通过从 OHC 调查数据($N=7\ 841$ 人)中抽取一个 SRSWOR 样本($n=1\ 960$ 人)并估算构造的 $D=30$ 个组群中慢性病患者的总和,来演示组群估算方法。在总体中,组群规模从最小 81 人变化到最大 517 人。组群慢性病患者的比例在 18% 到 39% 之间变化,整体上的比例为 29%。组群内患上慢性病(二分变量)与年龄(年份来表示)的相关系数为 0.08 到 0.55,整体上为 0.28。

在抽样阶段,组群被看成非设计类型。因而,组群样本规模在抽样设计中不是固定的,而是随机变化的。首先,计算了霍维茨-汤普森估算值。随后,使用式 6.2 中的模型辅助 GREG 估算公式,将辅助数据纳入估算程序中。辅助变量 z 的取值为所有 OHC 数据中的个人都登记了的年龄。数据中的所有人也构成了我们的目标总体。因此,在这一假设情形下,研究变量 y 的组群总和 $T_d(d=1, \cdots, D)$ 对于所有组群而言为已知。并且,它可以用于比较组群总和中。

从范例 6.2 中的简单模型(1b)($y_k=\beta\times z_k+\varepsilon_k$)得出所有组群同一的比率 $R=T/T_z=(7.778\times10^{-3})$。因此,构造于这一 P-模型之上的 GREG 估算值是间接类型的。在 SRSWOR 样本 $n=1\ 960$ 元素的基础上,比率 R 的一个估计值 $\hat{r}=\hat{t}_{HT}/\hat{t}_{zHT}=7.651\times10^{-3}$。其中,$\hat{t}_{HT}$($=225.3$)是研究变量 y 的总和 T 的 HT 估计值,$\hat{t}_{zHT}$($=294\ 357.5$)是辅助变量 z 的总和 T_z 的 HT 估计值。预测值由公式 $\hat{y}_k=\hat{r}\times z_k(k=1, \cdots, 7\ 841)$ 计算。估算公式的另外表达式在式6.21 中。那里,抽样权重为 $w_k=N/n=7\ 841/1\ 960=4.001$,$T_{dz}$ 为已知辅助变量 z 的组群总和,$\hat{t}_{dzHT}=\sum_{k\in s_d}w_kz_k$ 是相应的 HT 估计值。

$$\begin{aligned}\hat{t}_{dHT} &= \sum_{k\in s_d}w_ky_k = \frac{N}{n}\sum_{k\in s_d}y_k \\ \hat{t}_{dGREG\text{-}P} &= \sum_{k\in U_d}\hat{y}_k + \sum_{k\in s_d}w_k(y_k-\hat{y}_k) = \hat{t}_{dHT}+\hat{r}(T_{dz}-\hat{t}_{dzHT}),\end{aligned} \quad (6.21)$$

其中,s_d(含有 n_d 个元素)与 U_d(含有 N_d 个元素)分别是一对属于组群 d 的样

本与总体元素,$d=1,\cdots,D$。注意,相应的间接估算公式为 $\hat{t}_{dSYN\text{-}P}=\sum_{k\in U_d}\hat{y}_k=T_{dz}\times\hat{r}$。与 GREG 估算公式一样,是基于简单模型的。

在检验准确性时,我们使用估算值$\hat{t}_d$估算的标准误 $\mathrm{s.e}(\hat{t}_d)$以及离异系数的百分比 $\mathrm{c.v}(\hat{t}_d)=100\times\mathrm{s.e}(\hat{t}_d)/\hat{t}_d$。方差估算公式为,

$$\hat{v}_{srs}(\hat{t}_{dHT})=N^2\left(1-\frac{n}{N}\right)\left(\frac{1}{n}\right)p_d\,\hat{s}_{dy}^2\left(1+\frac{q_d}{\mathrm{c.v}_{dy}^2}\right),\ \text{and}$$
$$\hat{v}_{srs}(\hat{t}_{dGREG-P})=N^2\left(1-\frac{n}{N}\right)\left(\frac{1}{n}\right)p_d\,\hat{s}_{d\hat{e}}^2\left(1+\frac{q_d}{\mathrm{c.v}_{d\hat{e}}^2}\right),\tag{6.22}$$

其中,$p_d=n_d/n$,$q_d=1-p_d$,方差估算值为$\hat{s}_{dy}^2=\sum_{k\in s_d}(y_k-\bar{y}_d)^2/(n_d-1)$及$\hat{s}_{d\hat{e}}^2=\sum_{k\in s_d}(\hat{e}_k-\bar{\hat{e}}_d)^2/(n_d-1)$,估算的离异系数为 $\mathrm{c.v}_{dy}=\hat{s}_{dy}/\bar{y}_d$,$\mathrm{c.v}_{\hat{de}}=\hat{s}_{d\hat{e}}/\bar{\hat{e}}_d$,而$\bar{y}_d=\sum_{k\in s_d}y_k/n_d$,$\bar{\hat{e}}_d=\sum_{k\in s_d}\hat{e}_k/n_d$,残差 $\hat{e}_k=y_k-\hat{r}\times z_k$。

在选中的样本中,组群规模从 24 到 132,平均规模为 65。这样的情形可以使用基于设计的组群总和估算。我们首先看一下霍维茨-汤普森估算值$\hat{t}_{dHT}$与间接 GREG 估算值$\hat{t}_{dGREG\text{-}P}$的平均效果。在表 6.5 的第一部分,计算了三个组群规模登记中的相对差的绝对值的简单平均值$|\bar{\hat{t}}-\bar{T}|/\bar{T}$,其中的$\bar{\hat{t}}$是组群总和$\hat{t}_d$的均值,$\bar{T}$是给定规模等级的真实值 T_d 的均值。随着组群规模的增加,HT 与 GREG 估计值的相对差的绝对值倾向于降低。对于给定的规模等级,这两个数字几乎重合。对于 GREG 与 HT 估算值,选中的样本规模与离异系数有明显的联系:正如表 6.5 的第二部分中的平均离异系数所显示,组群规模增加,样本离异系数降低。总体上讲,GREG 估算值的离异系数要小些。

表 6.5　分组群样本规模的霍维茨-汤普森估计值平均绝对相对差与平均离异系数

规模分组	平均绝对相对差(%)		平均离异系数(%)	
	HT 估算值	GREG 估算值	HT 估算值	GREG 估算值
~39	10.6	10.2	30.8	24.7
40~79	2.0	3.4	23.5	19.8
80~	3.2	3.7	16.0	13.6
合计	1.8	1.7	23.0	19.0

在以组群规模排序的表 6.6 中,给出了 30 个组群离异系数的点估计、标准误与离异系数。与 HT 估算值$\hat{t}_{dHT}$相比,使用了辅助信息的模型辅助 GREG 估算值$\hat{t}_{dGREG\text{-}P}$显然提高了准确性。在所有 30 个组群中,估算的 GREG 标准误小于 HT 标准误。与预想的一致,在大多数组群中,估算的 GREG 离异系数也要小些。

表 6.6 OHC 数据中 SRSWOR 样本(n = 1 960)中慢性病患病人数估计值

组群				总和估计值		标准误		离异系数(%)	
d	n_d	N_d	T_d	$\hat{t}_{dHT}$	$\hat{t}_{dGREG}$	s.e($\hat{t}_{dHT}$)	s.e($\hat{t}_{dGREG}$)	c.v($\hat{t}_{dHT}$)	c.v($\hat{t}_{dGREG}$)
组群样本规模 $n_d<40$									
20	24	101	31	32.0	31.6	9.77	7.13	30.5	22.5
10	26	81	27	32.0	25.6	10.83	8.05	33.8	31.5
18	26	129	36	20.0	27.2	7.60	6.95	38.0	25.5
23	31	156	57	44.0	53.2	10.82	9.10	24.6	17.1
8	35	141	29	24.0	24.5	8.57	7.88	35.7	32.2
30	36	146	34	32.0	33.8	9.86	8.56	30.8	25.3
3	37	133	29	36.0	32.6	10.77	8.73	29.9	26.8
16	37	165	45	52.0	54.8	12.14	9.15	23.3	16.7
组群样本规模 $40\leq n_d<80$									
1	41	181	33	40.0	43.0	10.80	9.15	27.0	21.3
21	43	153	48	64.0	55.3	14.55	10.93	22.7	19.8
6	45	188	52	24.0	26.6	8.51	7.67	35.5	28.9
28	51	194	74	88.0	85.4	16.61	11.65	18.9	13.6
24	53	200	55	56.0	55.7	13.21	11.06	23.6	19.9
22	57	242	96	112.0	115.0	17.79	13.08	15.9	11.4
15	58	252	61	60.0	66.4	13.20	11.90	22.0	17.9
11	59	187	47	52.0	39.5	13.30	10.89	25.6	27.6
13	69	305	89	80.0	88.5	15.10	12.86	18.9	14.5
12	73	311	95	56.0	65.9	12.85	11.40	22.9	17.3
4	76	295	65	68.0	68.1	14.39	12.17	21.2	17.9
7	78	292	52	40.0	36.3	11.09	10.17	27.7	28.0
组群样本规模 $n_d\geq 80$									
2	84	352	86	76.0	78.6	14.95	13.49	19.7	17.2
5	86	323	66	76.0	70.5	15.31	13.62	20.1	19.3
26	89	364	124	124.0	126.0	19.07	15.72	15.4	12.5
29	90	365	128	124.0	124.5	19.12	15.10	15.4	12.1
25	91	339	114	112.0	101.6	18.68	14.81	16.7	14.6
17	99	426	139	176.0	183.3	22.11	16.72	12.6	9.1
9	103	366	89	88.0	79.3	16.66	13.82	18.9	17.4
19	115	490	165	152.0	160.0	20.81	17.13	13.7	10.7
14	116	447	130	136.0	128.4	20.31	16.28	14.9	12.7
27	132	517	197	176.0	173.8	22.94	17.51	13.0	10.1
合计	1 960	7 841	2 293	2 252.3	2 254.8	69.42	66.88	3.1	3.0

表中统计量为:组群样本规模 n_d、组群规模 N_d、总体总和 T_d、HT 与 GREG 估计值的点估计、标准误、离异系数(%)估计值(分组群样本规模)。

让我们以讨论在选中样本的情形下 GREG 估算值与模型决定的间接 SYN 估算值$\hat{t}_{dSYN\text{-}P} = T_{dz} \times \hat{r}$之间的关系来结束本例。从式 6.21 中 GREG 表达式，我们得到第一组群($n_1 = 41$)

$$\hat{t}_{1GREG\text{-}P} = \sum_{k \in U_1} \hat{y}_k + \sum_{k \in s_1} w_k (y_k - \hat{y}_k)$$

$$= 45.43 + 4.001 \times (-0.5974) = 43.04,$$

其中，第 1 组群的预测值$\hat{y}_k$的和的计算为$\sum_{k \in U_1} \hat{y}_k = T_{1z} \times \hat{r} = 5\,937 \times 0.007\,651\,5 = 45.43$。这是第 1 组群的合成估计值$\hat{t}_{1SYN\text{-}P}$。又如，对于组群 $d = 19$($n_{19} = 115$)，我们得到$\hat{t}_{19GREG\text{-}P} = 160.00$与$\hat{t}_{19SYN\text{-}P} = 138.09$，而真实值 $T_{19} = 165$。在这些组群中，GREG 估算公式中的偏差修正项正好成功地修正了 SYN 估算值的偏差。但，这在所有组群中并不都成立。事实上，在 30 个组群中的 17 个，GREG 估算值比 SYN 估算值更为成功。因为在几个组群中，其修正项方向正确但力度过大。在估算 SYN 估计值的准确度时，应当使用估算的均方差(MSE)，这是因为 SYN 并不是设计偏差的。在章节 6.4 中，以及本书的扩展网页里，我们进一步讨论 GREG 与 SYN 估算公式的关系。

6.4　估算公式的进一步比较

在本章节，我们使用蒙特卡洛模拟方法进一步审视组群总和的模型决定与模型辅助估算公式的特征。为了这一目的，我们仍然使用 OHC 调查数据。为了在经验上考察不同的 SYN 与 GREG 组群估算公式的理论特征(偏差与准确度)，我们有如下设定。第一，与范例 6.2 中相似，我们将 OHC 数据当成一个规模为 7 841 个元素的总体框架，并且在个体层次含有必要的辅助数据。第二，对于这总体框架数据，我们构造一个总数为 60 个组群的组群结构。这是因为，我们想要考察样本规模较小的组群。最后，从这个构造的假设的总体框架中，用 SRSWOR 的方法，以非设计组群结构的方式，抽取一大批规模为 1 000个元素的样本。通过计算这些模拟样本的平均估计值，我们来考察估算公式的偏差与准确度。

我们假定(根据资料 6.1 中的原则)，构造的 $N = 7\,841$ 人、$D = 60$ 个组群的 OHC 总体框架包含个体标签、组群标识变量、$n = 1\,000$ 元素的 SRSWOR 样本中所有 $k \in U$ 元素的选中概率，以及辅助 z-变量年龄的取值。样本元素的二分回应变量 y 为慢性病(取值 0:没有;取值 1:患有)。

根据线性模型的通用形式 $y_k = \beta_0 + u_{0d} + \beta_1 z_k + \varepsilon_k$，间接 SYN 与 GREG 估算公式使用 P-模型与 D-模型。在混合 D-模型的情形下，用限制性最大似然法(REML)与通用最小二乘法(GLS)来估算参数，并计算预测值$\hat{y}_k = \hat{\beta}_0 + \hat{u}_{0d} +$

$\hat{\beta}_1 z_k, k \in U$。在两个模型中，残差的计算是一样的，$\hat{e}_k = y_k - \hat{y}_k, k \in s$。将这些数据在个人层次与总体框合并（见表6.3）就可以用于组群估算了。

组群总和用以下公式来估算：

$$T_d = \sum_{k \in U_d} Y_k, \quad d = 1, \cdots, D。$$

间接估算公式为，

$$\hat{t}_{dSYN} = \sum_{k \in U_d} \hat{y}_k, d = 1, \cdots, D\text{（合成估算公式）},$$

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} w_k (y_k - \hat{y}_k), d = 1, \cdots, D\text{（通用回归估算公式）}。$$

这些公式中的预测值$\hat{y}_k$（$k \in U$）、观测的 y-数据 y_k、抽样权重 w_k 及残差$\hat{e}_k$（$k \in s$）提供了计算组群总和估计值的数据。间接估算值使用固定效应 P-模型与混合的 D-模型。对于合成估算值$\hat{t}_{dSYN\text{-}P}$与$\hat{t}_{dMSYN\text{-}D}$，仅使用$\hat{y}_k$ 预测值。对于 GREG 估算值$\hat{t}_{dGREG\text{-}P}$与$\hat{t}_{dMGREG\text{-}D}$，则使用了预测值$\hat{y}_k$、观测值 y_k、抽样权重 w_k，以及残差$\hat{e}_k = y_k - \hat{y}_k$。在这里讨论的 SRSWOR 情形，权重 $w_k = N/n$ 是常数，所有样本的残差和为 $\sum_{k \in s} \hat{e}_k = 0$。注意，这对于组群并不一定成立，因为我们是在使用组群总和的间接估算公式。

使用重复性的 K 个蒙特卡洛样本 s_v（$v = 1, 2, \cdots, K$）得到的估计值 $\hat{t}_d(s_v)$，我们比较各个估算值的偏差与准确度。对于每一个组群 $d = 1, \cdots, D$，计算了以下偏差与准确度的蒙特卡洛统计量。我们使用两种准确度的统计量，相对均方差的平方根（RRMSE）与绝对相对误差的中位值（MdARE）。这是因为对于二分变量而言，从这两个统计量中有时会得出不同的结论。

（ⅰ）绝对相对偏差（ARB），定义为偏差的绝对值与真实值之间的比率：

$$\left| \frac{1}{K} \sum_{v=1}^{K} \hat{t}_d(s_v) - T_d \right| / T_d。$$

（ⅱ）相对均方差的平方根（RRMSE），定义为均方差的平方根与真实值之间的比率：

$$\sqrt{\frac{1}{K} \sum_{v=1}^{K} [\hat{t}_d(s_v) - T_d]^2} / T_d。$$

（ⅲ）绝对相对误差的中位值（MdARE）的定义如下。对于每一个模拟样本 $s_v, v = 1, 2, \cdots, K$，得到其绝对相对误差，然后得出模拟的 K 个样本的中位值：

$$\underset{\text{over } v - 1, \cdots, K}{\text{Median}} \{ | \hat{t}_d(s_v) - T_d | / T_d \}。$$

表6.7给出了这一简单练习的实验设计特征小结。

表 6.7 蒙特卡洛模拟的技术细节小结

总体：	模型：	目标参数：
规模为 7 841 人的 OHC 调查总体框 样本规模：$n = 1\ 000$ 人 组群数：$D = 60$ 模拟样本数：$K = 500$ 独立的 SRSWOR 样本（非设计组群结构） 回应变量 y：慢性病 （二分：0—没病，1—患病） 辅助 z-数据：组群标识变量 年龄（年份表示）	(1a) 仅含截距的线性固定效应 P-模型： $y_k = \beta_0 + \varepsilon_k$ (1b) 以年龄为预测变量的固定效应 P-模型： $y_k = \beta_0 + \beta_1 z_k + \varepsilon_k$ (2a) 含有随机截距的线性混合 D-模型： $y_k = \beta_0 + u_{0d} + \varepsilon_k$ (2b) 以年龄为预测变量的混合 D-模型： $y_k = \beta_0 + u_{0d} + \beta_1 z_k + \varepsilon_k$	慢性病人数的组群总和 T_d，$d = 1, \cdots, 60$ 使用的估算值： SYN 估算值： 固定效应 P-模型的$\hat{t}_{dSYN\text{-}P}$ 两级线性 D-模型的$\hat{t}_{dMSYN\text{-}D}$ GREG 估算值： 固定效应 P-模型的$\hat{t}_{dGREG\text{-}P}$ 两级线性 D-模型的$\hat{t}_{dMGREG\text{-}D}$ 结果统计量： 平均计算所有组群规模层级： ARB 绝对相对偏差 RRMSE 相对均方差平方根 MdARE 绝对相对误差中位值

表 6.8 的 A 部分给出了简单模型(1a)与(1b)的结果小结，B 部分则给出了更为复杂的模型(1b)与(2b)的结果。结果显示，对于所有模型以及所有规模等级，由平均绝对相对偏差 ARB 来测量的、GREG 估算值 GREG-P 与 GREG-D 的偏差，都可以忽略不计。SYN 类型估算值的偏差随着模型选择变化。对于极端简单的固定效应 P-模型(1a)，SYN-P 的偏差相当大；当使用更符合实际的固定效应模型(1b)时，偏差减小。对于给出了最小偏差数字的 SYN 混合模型(2a)与(2b)，有着相似的结果。特别在较小组群中，对于所有模型类型以及对于 RRMSE 与 MdARE 统计量而言，SYN 估算值的准确度高于 GREG 估算值。但是，当组群样本规模增加时，准确度的差异就减小了。

表 6.8 中的结果还显示，SYN 类型估算值比 GREG 类型估算值的模型提升——从一个“较差”模型变成“较好”模型——更显著。注意，在这样的练习中，我们需要总体框架与样本数据在个人层次的合并。本书的扩展网页有这些数据。

表 6.8 不同模型选择下慢性病患病人数组群总和的 SYN 与 GREG 估计值的模拟结果（$K=500$ 的独立 SRSWOR 样本，每个含 $n=1\ 000$ 元素）

估计值	组群样本规模分组	组群间平均值					
		总体中组群总和	组群总和估计值	绝对相对偏差 ARB%	相对均方差平方根 RRMSE%	绝对相对误差中位值 MdARE%	组群样本规模
A. 固定效应 P-模型 $y_k=\beta_0+\varepsilon_k$ 与混合效应 D-模型 $y_k=\beta_0+\mu_{0d}+\varepsilon_k$。							
SYN-P	0～10	17.5	13.7	36.9	37.4	37.0	5.6
	11～20	37.0	34.4	50.3	50.7	50.3	14.1
	21～	62.4	78.8	43.6	44.2	43.6	32.4
	合计	38.2	41.2	43.5	44.0	43.5	16.9
MSYN-D	0～10	17.5	14.9	25.1	33.0	27.9	5.6
	11～20	37.0	35.7	22.7	33.3	25.0	14.1
	21～	62.4	66.3	11.6	26.0	17.4	32.4
	合计	38.2	38.2	20.0	30.9	23.6	16.9
GREG-P	0～10	17.5	17.5	2.4	55.2	39.5	5.6
	11～20	37.0	37.0	1.6	40.7	27.8	14.1
	21～	62.4	62.4	1.1	31.1	20.8	32.4
	合计	38.2	38.2	1.7	42.8	29.7	16.9
MGREG-D	0～10	17.5	17.3	2.6	53.5	38.9	5.6
	11～20	37.0	37.0	1.9	39.5	27.3	14.1
	21～	62.4	62.5	1.1	30.3	20.2	32.4
	合计	38.2	38.2	1.9	41.5	29.1	16.9
B. 固定效应 P-模型 $y_k=\beta_0+\beta_1 z_k+\varepsilon_k$ 与混合 D-模型 $y_k=\beta_0+\mu_{0d}+\beta_1 z_k+\varepsilon_k$。							
SYN-P	0～10	17.5	18.0	27.0	28.1	27.1	5.6
	11～20	37.0	36.6	19.6	20.8	19.7	14.1
	21～	62.4	62.0	12.1	13.9	12.5	32.4
	合计	38.2	38.1	19.8	21.2	20.0	16.9
MSYN-D	0～10	17.5	18.0	25.9	27.5	26.4	5.6
	11～20	37.0	36.6	17.7	20.2	18.5	14.1
	21～	62.4	62.1	9.7	14.4	11.6	32.4
	合计	38.2	38.2	18.1	20.9	19.1	16.9
GREG-P	0～10	17.5	17.5	2.7	53.0	38.5	5.6
	11～20	37.0	37.0	1.4	38.9	26.5	14.1
	21～	62.4	62.5	1.1	30.0	20.2	32.4
	合计	38.2	38.2	1.8	41.0	28.7	16.9
MGREG-D	0～10	17.5	17.5	2.7	52.8	38.4	5.6
	11～20	37.0	37.0	1.5	38.8	26.4	14.1
	21～	62.4	62.5	1.0	29.8	20.2	32.4
	合计	38.2	38.2	1.8	40.8	28.6	16.9

6.5 本章小结与更多的文献

在本章,我们集中讨论了组群的基于设计的模型辅助估算。这样的方法在生成官方统计数字中经常使用。在估算组群总和的过程中,我们有几个假定。特别地,在给定的统计基础上,我们假定有,总体框架登记、个人及加总层次的辅助数据、独特的标识可以让从样本调查数据与统计登记中得来的数据相合并。我们相信,满足这些条件可以为抽样设计与组群估算提供更多的灵活性。比如,如果需要的话,数据可以加总到更高的层次。单位层次的数据与建模对基于设计的模型辅助与模型决定的组群估算均有益处。看起来,细心与合乎实际的建模对于组群的模型决定估算特别重要。上述小规模的模拟显示了这一点。本章例子中的材料在扩展网页中有更为详细的讨论。

在实际中,基于设计的模型辅助估算更常用于组群样本规模较大的情形下。对于小组群而言,则使用小区域估算。对于组群估算,如果可能,推荐在抽样阶段将组群定义为层级,并使用一个恰当的配额方案,以便对于所有组群均获得一个较大规模的样本。在估算阶段,更好的方法是通过精心的模型选择,纳入强有力的辅助数据。

为了补充本章早些时候提到的文献,组群基于设计的模型辅助估算的文献有埃斯特瓦等(Estevao et al. ,1995)与埃斯特瓦与桑德尔(Estevao and Särndal,1999)。雷同能与维建能(Lehtonen and Veijanen,1998)讨论了诸如多值对数回归 GREG 估算的非线性 GREG 估算公式。

除了拉奥(Rao,2003)以外,小区域估算的模型决定方法在戈什与拉奥(Ghosh and Rao,1994)与拉奥(Rao,1999)中也有讲解。尤与拉奥(You and Rao,2002)讨论了设计调查权重的虚假 EBLUP 估算公式。基础的模型及其特征是其文献的重要主题(Ghosh et al. ,1998;Marker,1999;Moura and Holt,1999;Prasad and Rao,1999;Feder et al. ,2000)。也有许多从贝叶斯角度讨论小区域估算的文献,包括经验贝叶斯与多层级贝叶斯技术(Datta et al. ,1999;Ghosh and Natarajan,1999;You and Rao,1999)。近来的文献在小区域估算中,将概率频次恒定(frequentist)与贝叶斯方法相联系(Singh et al. ,1998)。瓦廉特等(Valliant et al. ,2000)用预测的方法讨论了小区域估算。

第7章 单维与二维表格分析

Analysis of One-way and Two-way Tables

在分析复杂调查中，经常会分析单维与二维频次表格。为了检验数据的拟合假设、同质性或是独立性，通常将已有的调查数据依据一个或是两个类别变量列表。例如，MFH 调查中 30 ~ 64 岁男性的年龄分布拟合度可以与相应的总体人口年龄分布相比照。又如，在 OHC 调查数据中，将应答者性别与二分回应变量 CHRON（慢性病）列联成 2 × 2 的表格，来检验男女两个性别的 CHRON 比例相同的原假设。另外，我们也可以检验回应变量 CHRON 与 PSYCH 类别变量——心理或是精神症状变量的第一个主成分取值中各个类别之间的独立性。在简单随机的情形下，可以根据标准的皮尔逊卡方统计量来做出这些检验的有效推论。但在更复杂的设计中，由于整群效应，检验过程更为复杂。

在简单随机抽样中，对于 $r \times c$ 频次列联表的同质与独立检验，皮尔逊检验统计量趋近于自由度为 $(r-1)(c-1)$ 的卡方。但是这一标准的渐进性特征，并不适用于基于整群抽样的复杂调查的频次表。列表变量间的正的群内相关系数，使得这一检验结果与较小的显著水平相比过大。因此，观测到的检验统计量过大而导致错误的推论。

复杂调查中的有效推论，需要对皮尔逊检验统计量做某些修正，如根据整群而自动调整拉奥-斯科特校正或是沃尔德检验统计量，章节 7.1 以一个简单拟合度的检验来演示这两种方法。章节 7.2 进一步讨论拟合度检验。章节 7.3 给出了二维表检验的基本内容。我们在章节 7.4 中，审视二维表格中同质假设的检验统计量，而在章节 7.5 中讨论两个类别变量的独立性。第 5 章中描述过的、涉及整群设计的 OHC 与 MFH 调查将用于各个例子中。

7.1 导入的例子

二分检验与有效样本规模

让我们先考虑一个使用于 OHC 调查的简单拟合度检验的假想例子。这一例子最早见于萨德曼(Sudman,1976),拉奥与托马斯(Rao and Thomas,1988)也演示了这一例子。从工业公司的大规模总体群中抽取一个 $m=50$ 的样本。我们假定,在每一个样本整群 $i=1, \cdots, 50$ 中,有 $n_i=20$ 个雇员。样本元素规模为 $n=1\ 000$。给定这样的抽样设计下恰当的数据,可以根据去年的情况,来检验职业健康保健(OHC)的覆盖率——未知的雇员总体中被职业健康服务覆盖的比例 p——是否为 80%。原假设可以写成为 $H_0:p=p_0=0.8$。令这一检验的限制性水平为 $\alpha=5\%$。

得到调查估计值 $\hat{p}=n_1/n=0.84$,其中 $n_1=840$ 是样本中享有 OH 服务的工人。我们选中二分检验,用标准正态分布 $N(0,1)$来表示,其大样本的检验统计量为

$$Z = |\hat{p}-p_0| / \sqrt{p_0(1-p_0)/n}, \tag{7.1}$$

其中的分母是在原假设下估计值$\hat{p}$的标准误。在计算 Z 的过程中,我们假定放回式简单随机抽样,以及使用考虑整群效应的基于设计的方法。在这个简单例子中,计算 Z 的观察值所需的$\hat{p}$的标准误在两种方法中均是建立在二分假设的基础之上,但是样本规模则并不相同。

在简单随机抽样假定下的检验中,我们忽略整群,并在标准误公式中使用实际上的样本规模 $n=1\ 000$。式 7.1 中的检验统计量的观测值为,

$$Z_{bin} = |\hat{p}-p_0| / \sqrt{p_0(1-p_0)/1\ 000} = 3.162 > Z_{0.025} = 1.96,$$

其中,$\sqrt{0.8(1-0.8)/1\ 000}=0.012\ 6$ 是$\hat{p}$的标准误。与标准 $N(0,1)$分布的相应的临界值相比较,这一结果显然意味着要拒绝原假设。

看起来,当 OHC 覆盖某一公司时,其中的每一个雇员均享有 OH 服务。这一重要的信息在前面的检验中被忽略了。事实上,从某一样本公司中选出多于一个以上的个人,并不能给我们更多的 OHC 覆盖率的信息。所以,与前面检验假定的 1 000 相比照,有效的样本规模为$\bar{n}=50$。回忆一下,有效样本规模是指一个简单随机样本规模,对于一个未知参数 p 的估计值,能够与实际中整群抽样设计得到的 $n=1\ 000$ 样本给出相同的精度。

使用有效样本规模,我们得到基于设计的检验,

$$Z_{des} = |\hat{p}-p_0| / \sqrt{p_0(1-p_0)/50} = 0.707,$$

其中,$\sqrt{0.8(1-0.8)/50}=0.056\ 6$ 远大于前面检验中的标准误。因而,观测

值 Z_{des} 要小于 Z_{bin}。现在的结果显示,不能拒绝原假设。接下来,我们研究更为普遍的例子,并引入另外的可以成功剔除掉整群效应的检验统计量。

皮尔逊检验统计量与拉奥-斯科特修正

与基于设计的 Z_{des} 相比,二分检验统计量 Z_{bin} 更为松散。这是因为,Z_{bin} 并没有考虑整群的因素。让我们通过构建一个相应的皮尔逊检验统计量 X_P^2 来更为详细地检视检验统计量 Z_{bin} 的渐进性趋势。为了这一目的,将要使用以下频次表,

j	n_j	p_{0j}
1	840	0.8
2	160	0.2
合计	1 000	1.0

其中,n_j 是方格频次的观测值,p_{0j} 为假定的方格比例。在有限总体的框架内,总体有 N 个元素,令未知方格比例为 $p_j = N_j / N$,而 N_j 为方格 j 中的总体元素。p_j 可以被看成是一个超总体框架下的未知方格概率。简单拟合度假设 $H_0: p_j = p_{0j}(j=1,2)$ 的皮尔逊检验统计量为,

$$X_P^2 = \sum_{j=1}^{2} (n_j - np_{0j})^2/(np_{0j}) = n\sum_{j=1}^{2} (\hat{p}_j - p_{0j})^2/p_{0j}, \qquad (7.2)$$

其中,比例 $\hat{p}_j = n_j/n$ 是 N_j 的样本值 n_j 的参数 p_j 的估计值。在两个方格的情况下,$\hat{p}_2 = 1 - \hat{p}_1$,因而,$p_{02} = 1 - p_{01}$,统计量 Z_{bin} 与 X_P^2 相似,

$$X_P^2 = n\sum_{j=1}^{2} (\hat{p}_j - p_{0j})^2/p_{0j} = (\hat{p} - p_0)^2/(p_0(1 - p_0)/n) = Z_{bin}^2,$$

其中,$\hat{p} = \hat{p}_1$,$p_0 = p_{01}$。在两个方格情况下,拟合度检验统计量 X_P^2 的自由度为 1,因为有一个限制条件(比例之和为 1),没有必要估计参数。

拉奥与斯科特(Rao and Scott,1981)给出了皮尔逊检验统计量 X_p^2 渐进分布更为普遍的结果。在两个方格的情况下,检验统计量 X_p^2 近似于随机变量 dW 的分布。W 是自由度为 1 的卡方随机变量 χ_1^2,而 d 表示比例估计值 $\hat{p}$ 的设计效应。可以通过 $d = V_{des}(\hat{p}) / V_{bin}(\hat{p})$ 得到设计效应,其中,$V_{des}(\hat{p}) = p_0(1 - p_0) / \bar{n}$ 是估计值 $\hat{p}$ 的设计方差,$\bar{n}$ 表示有效样本规模,$V_{bin}(\hat{p}) = p_0(1 - p_0) / n$ 是标准的二分方差。因而,在这一情形下,设计效应减小到 $d = n / \bar{n}$,亦意味着有效样本规模为 $\bar{n} = n / d$。

如果样本中的雇员实际上直接从雇员总体中以简单随机抽样抽取,我们有 $d = 1$,因为 V_{des} 与 V_{bin} 相等。在这种两个方格的情形下,皮尔逊检验统计量 X_P^2 趋近于一个自由度的卡方分布。但是如果样本实际上是整群抽样,群内正

相关造成设计效应 d 大于 1。由于此，统计量 X_P^2 并不趋近于恰当自由度的卡方分布。

考虑到群内正相关对于皮尔逊检验统计量 X_P^2 的渐进分布的影响，下一步将推导一个有效的检验方法。通常，X_P^2 的公式中无法将群内相关纳入其中，因而需要构造一个外在的 X_P^2 的相关系数。为了这一目的，首先注意到 X_P^2 的渐进期望值为 $E(X_P^2)=d$。在群内正相关的情形下，它大于通常期望值 1。由于 $E(X_P^2/d)=E(\chi_1^2)=1$，我们可以将检验统计量的观测值除以设计效应，构建一个简单 X_P^2 的拉奥-斯科特相关系数。调整整群效应的检验统计量为，

$$X_P^2(d) = X_P^2/d \tag{7.3}$$

它在两个方格的情形下，趋近于自由度为 1 的卡方分布。

对相应的似然比率(LR)，检验系数 X_{LR}^2 可以做一个类似的修正，它在两方格情形下的数值为，

$$X_{LR}^2 = 2n\sum_{j=1}^{2}\hat{p}_j\log(\hat{p}_j/p_{0j})。 \tag{7.4}$$

在简单随机抽样中，当原假设为真时，统计量 X_{LR}^2 也趋近于自由度为 1 的卡方分布。在整群设计中，相应的修正后的检验统计量为，

$$X_{LR}^2(d) = X_{LR}^2/d, \tag{7.5}$$

它近似于自由度为 1 的卡方分布。

接下来，我们计算 OHC 调查数据中修正过后的皮尔逊与 LR 检验统计量的大小。在修正中，我们需要设计效应，其大小为，

$$d = V_{des}(\hat{p})/V_{bin}(\hat{p}) = 0.003\,2/0.000\,16 = 20,$$

也可以作如下计算，$d=n/\overline{n}=1\,000/50=20$。

对于皮尔逊检验统计量，我们得到，

$$X_P^2 = (0.84-0.80)^2/(0.80\times 0.20/1\,000) = 10.00$$

其 p-值为 0.001 6。而拉奥-斯科特修正的皮尔逊检验统计量则为，

$$X_P^2(d) = X_P^2/d = Z_{bin}^2/d = 3.162^2/20 = 10.00/20 = 0.50,$$

其 p-值为 0.479 5。应当注意到，$Z_{des}^2=0.707^2=0.50$。即，与期望相同，$Z_{des}^2=X_P^2(d)$。对于似然检验统计量以及相应的拉奥-斯科特修正，我们得到，

$$X_{LR}^2 = 2\times 1\,000\times[0.84\times\log(0.84/0.80)+ 0.16\times\log(0.16/0.20)] = 10.56,$$

其 p-值为 0.001 2，且，

$$X_{LR}^2(d) = X_{LR}^2/d = 10.560/20 = 0.528$$

其 p-值为 0.467 5。

由于群内正相关是完整的，所以观测到的设计效应 $d=20$ 较为异常的大。从等式 $d=1+(\overline{m}-1)\rho_{int}$ 计算得出，群内相关系数为 $\rho_{int}=1$，其中的 $\overline{m}=20$ 是平均整群规模。在实际中，群内相关通常为正，但小于 1，设计效应的估计值 $\hat{d}$

相应的大于1。典型的$\hat{d}$小于3,对应的正的相关系数估计值为$\rho_{int} < 0.1$,其中$\overline{m}=20$。

内曼与沃尔德检验统计量

作为皮尔逊检验统计量的替代,可以计算简单拟合度假设的内曼检验统计量X_N^2。在两个方格的情形下,它简化为,

$$X_N^2 = n\sum_{j=1}^{2}(\hat{p}_j - p_{0j})^2/\hat{p}_j = (\hat{p} - p_0)^2/[\hat{p}(1-\hat{p})/n], \qquad (7.6)$$

它与皮尔逊检验统计量不同。因为,在分母中的假定比例p_{0j}由估计值$\hat{p}_j$所代替。在简单随机抽样的两个方格情形下,内曼检验统计量接近于自由度为1的卡方分布。但在整群抽样中,内曼检验统计量应当作与皮尔逊检验统计量相似的修正。内曼检验统计量的拉奥-斯科特修正为,

$$X_N^2(\hat{d}) = X_N^2/\hat{d} = \hat{d}^{-1}(\hat{p} - p_0)^2/[\hat{p}(1-\hat{p})/n]。 \qquad (7.7)$$

估算的设计效应由公式$\hat{d}=\hat{v}_{des}(\hat{p})/\hat{v}_{bin}(\hat{p})$计算得出。其中,$\hat{v}_{des}$是与实际抽样设计相对应的基于设计的$\hat{p}$的方差估计值,$\hat{v}_{bin}$则是二分分布中相应的方差估计值。

接下来,我们计算内曼检验统计量及其拉奥-斯科特修正。为了这一目标,我们使用估算的设计效应。首先得出基于设计的$\hat{p}$的方差估计值:

$$\begin{aligned}\hat{v}_{des}(\hat{p}) &= \sum_{i=1}^{m}(\hat{p}_i - \hat{p})^2/[m(m-1)] \\ &= \sum_{i=1}^{50}(\hat{p}_i - 0.84)^2/(50\times 49) \\ &= 0.002\,743,\end{aligned}$$

其中,m是样本整群数目,$\hat{p}_i$是样本整群i中OHC的覆盖率,$\hat{p}$则是整个样本中的估计值。应当注意到,$\hat{p}_i$取值0或1。可以使用一个二分方差估计值,

$$\hat{v}_{bin}(\hat{p}) = \hat{p}(1-\hat{p})/n = 0.000\,134,$$

计算出一个设计效应估计值。其大小为$\hat{d}=0.002\,743/0.000\,134=20.4$。另外,设计效应也可以估算为$\hat{d}=\hat{v}_{des}(\hat{p})/\hat{V}_{bin}(\hat{p})=17.1$。

内曼检验统计量的观测值为,

$$X_N^2 = (0.84-0.80)^2/(0.84\times 0.16/1\,000) = 11.90$$

其p-值为0.000 6。对于内曼检验统计量的拉奥-斯科特修正值,我们得到,

$$X_N^2(\hat{d}) = X_N^2/\hat{d} = 11.9/20.4 = 0.583$$

其p-值为0.445 1。注意,内曼检验统计量观测值及其拉奥-斯科特修正略大于皮尔逊统计量及其拉奥-斯科特修正。

内曼检验统计量X_N^2是沃尔德(Wald,1943)拟合度检验统计量的特殊形

式。与皮尔逊、似然以及内曼统计量不同，沃尔德统计量自动考虑到了群内相关。这一点可以从基于设计的沃尔德统计量的公式看出，其在两个方格的情形下简化为，

$$X_{des}^2 = (\hat{p} - p_0)^2 / \hat{v}_{des}, \tag{7.8}$$

其中，$\hat{v}_{des}$是$\hat{p}$基于设计的方差估计值。统计量 X_{des}^2在整群抽样设计中，并不借助于任何辅助相关而近似于自由度为 1 的卡方分布。在简单随机抽样中，式 7.8 中的$\hat{v}_{des}$由$\hat{v}_{bin}$替代，内曼统计量 X_{N}^2 与 X_{bin}^2所表示的沃尔德统计量相同。显然，在整群抽样中，X_{bin}^2也需要与内曼统计量相似的修正。

在计算基于设计的沃尔德统计量数值时，我们得到，

$$X_{des}^2 = (0.84 - 0.80)^2/0.002\ 743 = 0.583,$$

正如所期望的，它与拉奥-斯科特修正的内曼统计量相等。这显示了沃尔德统计量的灵活性。使用反应抽样设计复杂性的合适的方差估计值，我们不借助任何辅助修正，可以得到一个近似的有效的检验统计量。这可以看成是相对于拉奥-斯科特修正统计量的优势。但是，我们将在后面多于两个方格的更普遍的例子中看到，基于设计的沃尔德统计量有着特定的缺点。这是由某些小样本情形下方差估计值的潜在的不稳定性所造成的。

最后，我们在下面给出式 7.2 到式 7.8 检验统计量的检验结果，如下表。

检验统计量	df	观测值	p-值
皮尔逊			
X_{P}^2	1	10.00	0.001 6
$X_{P}^2(d)$（修正）	1	0.500	0.479 5
似然比率			
X_{LR}^2	1	10.56	0.001 2
$X_{LR}^2(d)$（修正）	1	0.528	0.467 5
内曼			
$X_{N}^2(=X_{bin}^2)$	1	11.90	0.000 6
$X_{N}^2(\hat{d})$（修正）	1	0.583	0.445 1
沃尔德			
X_{des}^2	1	0.583	0.445 1

这个例子中显示了两个考虑整群效应的检验统计量的主要方法。它们是用于皮尔逊、似然及内曼检验统计量的拉奥-斯科特修正方法与基于设计的沃尔德统计量。它们可以用于更为普遍的一维表格以及行列数大于 2 的二维表格。下面，我们讨论一个简单拟合度检验更普遍的例子，并给出其他统计量的

细节。然后，讨论二维表格的同质性与独立性假设检验。在检验方法中，我们将集中关注基于设计的沃尔德统计量以及各种皮尔逊与内曼检验统计量的拉奥-斯科特修正。

7.2 简单拟合度检验

与简单的两个方格的例子相比，多于两个方格情形下拟合度假设检验的有效方法要复杂得多。对于基于设计的沃尔德统计量以及皮尔逊与内曼的拉奥-斯科特修正，均是如此。我们接下来详细讨论这些检验方法。

由于基于设计的沃尔德统计量在复杂调查中的渐进性意义上是普遍正确的，所以它对于简单的拟合度假设给出了一个自然的检验方法。当有大量的样本整群时——OHC 调查正是这样的，沃尔德统计量在实际中足够有效。当样本整群数较小时，这一检验统计量可能造成不稳定的问题。这种情形下，将得到过大的统计量的观测值的结果。幸运的是，检验统计量的不稳定性效应能够为 *F*-校正所降低。另一个渐进性意义上普遍有效的检验方法，是皮尔逊与内曼检验统计量基于二阶的拉奥-斯科特修正。在这两个检验方法中，得到完整的基于设计的协方差矩阵是非常重要的。这需要元素层次的数据。

我们在实际中可能遇到无法得到元素层次的数据。比如，在对已公开的表格的二手分析中，几乎没有提供完整的基于设计的协方差矩阵的估计值。所以，不能使用沃尔德统计量或是二阶的拉奥-斯科特修正。但是，如果有恰当的设计效应估计值，特定的近似一阶修正是可能的。虽然，基于这些设计效应估计值的修正仅仅在特殊条件下才有效，但在很多情形下，与不正确的皮尔逊与内曼检验统计量相比，它们是更优的选择。

对于 $u \geqslant 2$ 方格而言，拟合度假设可以写成，$H_0: p_j = p_{0j}, j = 1, \cdots, u$，其中，$p_j = N_j/N$ 是未知方格比例，而 p_{0j} 是假定的方格比例。使用相应的向量，原假设可以轻易地改写成，$H_0: \mathbf{p} = \mathbf{p}_0$，其中 $\mathbf{p} = (p_1, \cdots, p_{u-1})'$ 是未知方格比例向量，$\mathbf{p}_0 = (p_{01}, \cdots, p_{0,u-1})'$ 是假定的方格比例向量。基于 n 个元素样本的方格比例的一致估计向量由 $\hat{\mathbf{p}} = (\hat{p}_1, \cdots, \hat{p}_{u-1})'$ 来表示。其中 $\hat{p}_j = \hat{n}_j/n$。$\hat{n}_j$ 是考虑个体元素选中概率不等并针对无应答修正的、换算加权过的方格频次，其中 $\sum_{j=1}^{u} \hat{n}_j = n$（见第 5 章）。当 n 并非预先确定时，通常针对总体子群时（正如在此假定的），$\hat{p}_j$ 是比率估计值。注意，在向量 $\mathbf{p}, \mathbf{p}_0, \hat{\mathbf{p}}$ 中，仅含有 $u-1$ 个元素。这是因为，各个比例的限制条件是合计为 1，如，$\hat{p}_u = 1 - \sum_{j=1}^{u-1} \hat{p}_j$。

基于设计的沃尔德统计量

简单拟合度假设的基于设计的沃尔德统计量 X^2_{des} 是作为调整过的皮尔逊

统计量的替代，在整群抽样设计情形下的两个方格例子中引入的。在多于两个方格的情形下，拟合度基于设计的沃尔德统计量略微复杂些：

$$X_{des}^2 = (\hat{\mathbf{p}} - \mathbf{p}_0)' \hat{\mathbf{V}}_{des}^{-1} (\hat{\mathbf{p}} - \mathbf{p}_0), \tag{7.9}$$

其中，$\hat{\mathbf{V}}_{des}$表示比例估计值向量$\hat{\mathbf{p}}$的真实协方差 $\mathbf{V}/n$ 的协方差矩阵的一致估计值。使用线性化方法可以得到$\hat{\mathbf{V}}_{des}$的估计值，也可以使用样本再使用方法，如折刀方法。当原假设为真时，统计量 X_{des}^2趋近于自由度为 $u-1$ 的卡方分布。这样就提供了一个复杂调查的检验方法。在实际中，当样本整群数目巨大而方格数目较小时，X_{des}^2是较为合理的。这种情形下，$\hat{\mathbf{V}}_{des}$的估计值也较为稳定。注意，式 7.8 中的统计量是式 7.9 中统计量的特例。

不稳定的情形

当样本整群数目 m 较小时，$\hat{\mathbf{V}}_{des}$可能遇到不稳定的问题。这是因为，估计值的自由度$f = m - H$ 较小。沃尔德统计量 X_{des}^2的不稳定估计值$\hat{\mathbf{V}}_{des}$的后果可能很严重，将使得统计量过大。在克服不稳定问题中，使用最为广泛的方法之一就是调整沃尔德统计量的自由度，得出一个设定为 F-分布的新的统计量。有两种 F-校正的沃尔德统计量。第一个为，

$$F_{1.des} = \frac{f-u+2}{f(u-1)} X_{des}^2, \tag{7.10}$$

它被看成是自由度为 $u-1$ 与$f-u+2$ 的 F-分布的随机变量。第二个为，

$$F_{2.des} = X_{des}^2/(u-1), \tag{7.11}$$

它是自由度为 $u-1$ 与f的 F-分布的。注意，当 $u=2$ 时，两个修正均得出原来的统计量。在两个方格的例子中，可以很容易地看出对 X_{des}^2的 F-校正的效果。当f较小时，X_{des}^2自由度为 1 与f的 F-分布的 p-值，大于自由度为 1 的卡方分布的 p-值。但，当f增加时，这一差异就消失了。因此，当f较大时，修正的效果并不明显。但对于较小的f，这样可以纠正并不准确的沃尔德统计量较大的取值；对于 $u > 2$，也是这样。

根据模拟，托马斯与拉奥（Thomas and Rao，1987）给出了在不稳定情形下，一个简单拟合度的各种检验统计量的结果。虽然，他们注意到了，与其他统计量相比，F-校正的沃尔德统计量 $F_{1.des}$并不全面占优。但是，在不稳定性不是异常严重的标准情形下，它的结果相对较好。在实际中，F-校正的沃尔德统计量使用广泛，并被用于调查分析的计算机软件中。

皮尔逊检验统计量与拉奥-斯科特修正

正如导入例子中指出的，基于简单随机抽样假设的检验统计量，需要修正

其整群效应，以满足相应的渐进性特征。让我们首先考虑皮尔逊检验统计量 X_P^2。这一统计量可以简洁地写成矩阵的形式：

$$X_P^2 = n\sum_{j=1}^{u}(\hat{p}_j - p_{0j})^2/p_{0j} = n(\hat{\mathbf{p}} - \mathbf{p}_0)'\mathbf{P}_0^{-1}(\hat{\mathbf{p}} - \mathbf{p}_0), \qquad (7.12)$$

其中，$\mathbf{P}_0 = \text{diag}(\mathbf{p}_0) - \mathbf{p}_0\mathbf{p}'_0$，$\mathbf{P}_0/n$ 是在原假设条件下 $\hat{\mathbf{p}}$ 的 $(u-1)\times(u-1)$ 维协方差矩阵，算符 diag($\mathbf{p}_0$)生成一个对角元素为 p_{0j} 的对角矩阵。协方差矩阵 $\mathbf{P}_0/n$ 是从 $u=2$ 到大于 2 个方格的扩展。注意，X_P^2 的矩阵形式的公式与式 7.9 中的沃尔德统计量相似，仅有的区别是使用了 $\mathbf{P}_0/n$ 而非 $\hat{\mathbf{V}}_{des}$。在两个方格的例子中，X_P^2 简化为前面提及的简单公式 $X_P^2 = (\hat{p}_1 - p_{01})^2/[p_{01}(1-p_{01})/n]$。其分母对应于在原假设条件下的二分方差。

为了检验皮尔逊检验统计量 X_P^2 的渐进性分布，我们将前面的两个方格情形下的结果推广到 $u > 2$ 的情形。在这种情形下，X_P^2 的渐进性分布为，$u-1$ 个自由度为 1 的独立的卡方随机变量 W_j 的加权和 $\delta_1 W_1 + \delta_2 W_2 + \cdots + \delta_{u-1}W_{u-1}$。权重 δ_j 是通用设计效应矩阵 $\mathbf{D} = \mathbf{P}_0^{-1}\mathbf{V}$ 的特征值。其中，$\mathbf{V}/n$ 是基于实际抽样设计的比例估计向量 $\hat{\mathbf{p}}$ 的真实的协方差矩阵。这些特征值也被称为通用设计效应。注意，通常意义上，它们与设计效应 d_j 并不相同。

当实际抽样设计为简单随机抽样时，通用设计效应 δ_j 均为 1。因为，真实的与假定的方差矩阵 $\mathbf{V}/n$ 与 $\mathbf{P}_0/n$ 相同。因此，通用设计效应矩阵成为一个单位矩阵。加权和 $\sum_{j=1}^{u-1}\delta_j W_j$ 简化为 $\sum_{j=1}^{u-1}W_j$，即是 $u-1$ 个独立的卡方随机变量 χ_1^2 之和，其分布显然是一个自由度为 $u-1$ 的 χ^2 分布。因此，在简单随机抽样下，皮尔逊统计量 X_P^2 是一个自由度为 $u-1$ 的渐进性的卡方分布。

当实际抽样设计涉及整群且更加复杂时，真实的 $\mathbf{V}/n$ 与假定的 $\mathbf{P}_0/n$ 并不一定相同。在这种情形下，通用设计效应 δ_j 并不等于 1。由于整群效应，δ_j 通常倾向于大于 1。因此，随机变量 $\sum_{j=1}^{u-1}\delta_j W_j$ 的渐进性分布并不假定为自由度为 $u-1$ 的卡方分布。所以，皮尔逊统计量需要与两个方格情形下相似的校正。但是，现在有了更多的修正皮尔逊统计量的方法，它们是拉奥与斯科特(Rao and Scott, 1981)提出的一阶与二阶拉奥-斯科特修正。一阶修正的目标是校正皮尔逊统计量的渐进性期望值，而二阶修正涉及渐进性地校正方差。技术上而言，两种方法都是基于估计的通用设计效应矩阵 $\hat{\mathbf{D}}$ 的特征值。

我们首先讨论简单的费勒基(Fellegi, 1980)与霍尔特等(Holt et al., 1981)提出的设计效应均值修正与一阶拉奥-斯科特修正。这些方法使用于无法得到整个基于设计的估计值 $\hat{\mathbf{V}}_{des}$ 的情形。当给出了这一估计值，则应当使用更加准确的二阶修正。

设计效应均值修正是基于比例 $\hat{p}_j$ 的估算的设计效应 $\hat{d}_j$。用皮尔逊统计量

的观测值除以平均设计效应,得到式 7.12 的修正统计量:

$$X_{\mathrm{P}}^{2}(\hat{d}_{\bullet}) = X_{\mathrm{P}}^{2}/\hat{d}_{\bullet}, \tag{7.13}$$

其中,$\hat{d}_{\bullet} = \sum_{j=1}^{u}\hat{d}_j/u$ 是未知设计效应 d_j 的均值 $\bar{d}$ 的估计值。我们用 $\hat{d}_j = \hat{v}_{des}(\hat{p}_j)/(\hat{p}_j(1-\hat{p}_j)/n)$ 来估算设计效应。其中,$\hat{v}_{des}(\hat{p}_j)$ 是比例估计值 $\hat{p}_j$ 的基于设计的方差估计值。因此,这一修正需要 u 个方格比例估计值的设计效应估计值。正的群内相关给出了均值 $\hat{d}_{\bullet}$ 大于 1。因而,deff 均值修正倾向于剔除 X_{P}^{2} 过宽的取值范围,通过计算有效样本 $\bar{n} = n/\hat{d}_{\bullet}$,并用 $\bar{n}$ 替代 X_{P}^{2}(式 7.12)中的 n,计算得出。

设计效应均值修正是一个近似。因而,它并不涉及对 X_{P}^{2} 的渐进期望值的准确修正。这是因为,设计效应的均值并不等于通用设计效应的均值。在原假设下,X_{P}^{2} 的渐进期望值 $E(X_{\mathrm{P}}^{2}) = \sum_{j=1}^{u-1}\delta_j$,因此,$E(X_{\mathrm{P}}^{2}/\bar{\delta}) = E(\chi_{u-1}^{2}) = u-1$。其中,特征值的均值为 $\bar{\delta} = \sum_{j=1}^{u-1}\delta_j/(u-1)$。这推导出了 X_{P}^{2} 的一阶拉奥 - 斯科特修正,

$$X_{\mathrm{P}}^{2}(\hat{\delta}_{\bullet}) = X_{\mathrm{P}}^{2}/\hat{\delta}_{\bullet}, \tag{7.14}$$

其中,$\hat{\delta}_{\bullet}$ 是未知特征值均值 $\bar{\delta}$ 的估计值。这一均值可以在不估算特征值的情形下,使用以下设计效应估计等式来计算,

$$(u-1)\hat{\delta}_{\bullet} = \sum_{j=1}^{u}\frac{\hat{p}_j}{p_{0j}}(1-\hat{p}_j)\hat{d}_j$$

另外的,$\hat{\delta}_{\bullet}$ 也可以通过利用等式 $\hat{\delta}_{\bullet} = \mathrm{tr}(\hat{\mathbf{D}})/(u-1)$,从通用设计效应矩阵估计值 $\hat{\mathbf{D}} = n\mathbf{P}_0^{-1}\hat{\mathbf{V}}_{des}$ 中得出,即,用矩阵 $\hat{\mathbf{D}}$ 的迹除以自由度。只有在所有的特征值 δ_j 相等时,修正的 $X_{\mathrm{P}}^{2}(\hat{\delta}_{\bullet})$ 是自由度为 $u-1$ 的渐进卡方分布。但是,在实际中当估算特征值 $\hat{\delta}_j$ 较小时,这一统计量也较为合理。因为,这一统计量仅仅需要设计效应估计值 $\hat{p}_j$,它也适用于给出设计效应估计值的已公开表格的二手分析。与设计效应均值修正 $X_{\mathrm{P}}^{2}(\hat{d}_{\bullet})$ 相比,一阶拉奥-斯科特修正 $X_{\mathrm{P}}^{2}(\hat{\delta}_{\bullet})$ 更为准确。前者可以看成是后者的保守替代。

式 7.14 中的一阶拉奥-斯科特修正在于改进皮尔逊检验统计量 X_{P}^{2},使得其渐进期望值等于自由度。当估计的特征值 $\hat{\delta}_j$ 较大时,则需要修正 X_{P}^{2} 的方差。使用萨特斯韦特(Satterthwaite,1946)方法的二阶拉奥-斯科特修正可以达到这一目标。皮尔逊统计量的二阶拉奥-斯科特修正为,

$$X_{\mathrm{P}}^{2}(\hat{\delta}_{\bullet},\hat{a}^{2}) = X_{\mathrm{P}}^{2}(\hat{\delta}_{\bullet})/(1+\hat{a}^{2}), \tag{7.15}$$

其中,未知特征值 δ_j 的离异系数平方 a^2 的估算公式为,

$$\hat{a}^2 = \sum_{j=1}^{u-1} \hat{\delta}_j^2 / ((u-1)\hat{\delta}_{\bullet}^2) - 1。$$

特征值的平方和的估算公式为，

$$\sum_{j=1}^{u-1} \hat{\delta}_j^2 = \mathrm{tr}(\hat{\mathbf{D}}^2) = n^2 \sum_{j=1}^{u} \sum_{k=1}^{u} \hat{v}_{des}^2(\hat{p}_j, \hat{p}_k) / p_{0j} p_{0k},$$

其中，$\hat{v}_{des}(\hat{p}_j, \hat{p}_k)$是 $\hat{p}_j$ 与$\hat{p}_k$ 的方差与协方差估计值。这一统计量的自由度也必须调整，$X_P^2(\hat{\delta}_{\bullet}, \hat{a}^2)$是萨特斯韦特修正自由度为 $df_S = (u-1)/(1+\hat{a}^2)$的渐进卡方分布。注意，在二阶修正中，需要完整的协方差矩阵估计值$\hat{\mathbf{V}}_{des}$；而在一阶修正中，仅需要方差估计值$\hat{v}_{des}$。

在不稳定的情形下，一阶拉奥-斯科特修正的 F-校正是有益的。其公式为，

$$FX_P^2(\hat{\delta}_{\bullet}) = X_P^2 / ((u-1)\hat{\delta}_{\bullet})。 \tag{7.16}$$

这一统计量是自由度为 $u-1$ 与f 的 F-分布。托马斯与拉奥（Thomas and Rao，1980）指出了这一统计量在不稳定情形下优于未改进的一阶修正。

内曼（多值沃尔德）统计量

前面，内曼检验统计量 X_N^2 被用作皮尔逊统计量的替代。内曼统计量对应于$\hat{\mathbf{p}}$被假定为多项分布而推导出的沃尔德统计量。内曼统计量为，

$$X_N^2 = n \sum_{j=1}^{u} (\hat{p}_j - p_{0j})^2 / \hat{p}_j = n(\hat{\mathbf{p}} - \mathbf{p}_0)' \hat{\mathbf{P}}^{-1} (\hat{\mathbf{p}} - \mathbf{p}_0), \tag{7.17}$$

其中，$\hat{\mathbf{P}} = \mathrm{diag}(\hat{\mathbf{p}}) - \hat{\mathbf{p}}\hat{\mathbf{p}}'$，而$\hat{\mathbf{P}}/n$ 是估计（实际）的多项协方差矩阵。注意，这一等式与式 7.9 与 7.12 的基于设计的沃尔德统计量与皮尔逊统计量相似，仅有的区别在于$\hat{\mathbf{P}}/n$ 代替了$\hat{\mathbf{V}}_{des}$或是 $\mathbf{P}_0/n$。在简单随机抽样下，X_N^2 是自由度为 $u-1$的渐进卡方分布。但在更为复杂的抽样设计下，这一统计量需要作与皮尔逊统计量相似的修正。所以，X_N^2 的设计效应均值修正为 $X_N^2(\hat{d}_{\bullet}) = X_N^2 / \hat{d}_{\bullet}$；一阶拉奥-斯科特修正为 $X_N^2(\hat{\delta}_{\bullet}) = X_N^2 / \hat{\delta}_{\bullet}$；二阶拉奥-斯科特修正为 $X_N^2(\hat{\delta}_{\bullet}, \hat{a}^2) = X_N^2(\hat{\delta}_{\bullet}) / (1+\hat{a}^2)$；以及一阶拉奥-斯科特修正的 F-校正为 $FX_N^2(\hat{\delta}_{\bullet}) = X_N^2(\hat{\delta}_{\bullet}) / (u-1)$。

检验统计量与分布特征

到目前为止，我们的讨论显示，检验统计量的渐进特征取决于统计量具体的抽样设计假设及其实际的抽样设计。更具体而言，令 $\mathbf{D} = \mathbf{P}^{-1}\mathbf{V}$ 为设计效应矩阵。其中的 $\mathbf{P}/n$ 是对应于假定抽样设计的协方差矩阵，而 $\mathbf{V}/n$ 是基于实际抽样设计的真实的协方差矩阵。检验统计量的渐进分布取决于这样的设计效应矩阵的特征值。如果所有的特征值等于 1，拟合度的检验统计量符合自由

度为 $u-1$ 的渐进卡方分布。

对于皮尔逊检验统计量而言,假定的协方差矩阵 $\mathbf{P}/n$ 是一个多项的$\mathbf{P}_0/n$。当实际设计为简单随机抽样时,真实的 $\mathbf{V}/n$ 与假定的 $\mathbf{P}/n$ 相等同,所有的特征值均为 1。但实际设计更为复杂时,协方差矩阵并不相等同,特征值也不等于 1。因此,需要对 X_P^2 进行修正。

对于基于设计的沃尔德统计量,因为假定与实际的抽样设计相同,情形又有区别。根据定义,$\mathbf{D}$ 中的 $\mathbf{V}/n$ 与 $\mathbf{P}/n$ 相同。因此,当实际设计为简单随机抽样时,我们有 $\mathbf{P}/n=\mathbf{V}/n=\mathbf{P}_0/n$;当实际设计涉及到整群与分层而更为复杂时,我们有 $\mathbf{P}/n=\mathbf{V}/n$。在两种情形下,对应的设计效应矩阵的特征值等于 1,不需要对 X_{des}^2进行修正。

残差分析

如果拟合度检验并不支持原假设,可以进行残差分析来检查相对于 H_0 的偏离。对于简单随机抽样,标准化残差的形式为,

$$\hat{e}_j=(\hat{p}_j-p_{0j})/\text{s. e}_{srs}(\hat{p}_j),j=1,\cdots,u, \tag{7.18}$$

其中,$\text{s. e}_{srs}(\hat{p}_j)$是所对应的多项协方差估计值$\hat{\mathbf{P}}/n$ 的对角元素的平方根。$\hat{e}_j$较大的绝对值意味着相当于 H_0 的偏离。但在复杂调查中,由于多项标准误倾向于低估真实的标准误,这些标准化残差可能较大。因此,我们使用相应的基于设计的标准误 $\text{s. e}_{des}(\hat{p}_j)$来推出基于设计的标准化残差。我们有,

$$\hat{e}_j=(\hat{p}_j-p_{0j})/\text{s. e}_{des}(\hat{p}_j),j=1,\cdots,u。 \tag{7.19}$$

显然,当设计效应明显大于 1 时,相对于多项标准误,式 7.19 得到较小的标准化残差。在原假设下,基于设计的标准化残差可以被当成近似的标准化正态变量,而它们可以与 $N(0,1)$分布的临界值相对照。

范例 7.1

MFH 调查中年龄分布的拟合度检验。我们考虑 MFH 调查中 30 ~ 64 岁男性子集相较于总体人口年龄分布的拟合度检验。我们之所以选择 MFH 设计的另一个目的,是为了演示小数目样本整群($m=48$)对于检验结果的影响。表 7.1 给出了样本及总体人口年龄分布,以及估算的比例估计值的方格设计效应。表中也包括了标准化的基于设计的残差。

表 7.1 MFH 调查中 30～64 岁男性的年龄组的估算与假设的年龄分布、年龄比例的设计效应估计值及标准残差

年 龄	n_j	估算值$\hat{p}_j$	假设值 p_{0j}	Deff $\hat{d}_j$	残差$\hat{e}_j$
30～44	1 329	0.492	0.521	1.51	-2.45
45～54	774	0.287	0.277	1.70	0.88
55～64	596	0.221	0.202	0.43	3.64
总体样本	2 699	1.000	1.000		

由于方格比例限制条件合计为 1，所以检验中自由度为 $u-1=2$。原假设为，$H_0:p_j=p_{0j}, j=1, 2, 3$。使用样本与总体比例$\hat{p}_j$ 与 p_{0j}以及设计效应估计值$\hat{d}_j$，从表 7.1 中可以计算出未修正的皮尔逊与内曼检验统计量、对于皮尔逊统计量的设计效应均值修正值与一阶拉奥-斯科特修正值。但是，二阶拉奥-斯科特修正值与沃尔德统计量需要完整的比例估计值$\hat{\mathbf{V}}_{des}$。可以通过线性化方法来获得这一估计值。为了信息的完整性，我们给出完整的 3 × 3 协方差矩阵估计值，

$$\hat{\mathbf{V}}_{des} = 10^{-5} \times \begin{bmatrix} 13.948\ 1 & -12.073\ 1 & -1.875\ 0 \\ -12.073\ 1 & 12.915\ 8 & -0.842\ 7 \\ -1.875\ 0 & -0.842\ 7 & 2.717\ 7 \end{bmatrix}。$$

为了比较，我们也给出多项式协方差矩阵 $\mathbf{P}_0/n=(\mathrm{diag}(\mathbf{p}_0)-\mathbf{p}_0\mathbf{p}_0')/2\ 699$，以及$\hat{\mathbf{P}}/n=(\mathrm{diag}(\hat{\mathbf{p}})-\hat{\mathbf{p}}\hat{\mathbf{p}}')/2\ 699$。它们是，

$$\mathbf{P}_0/n = 10^{-5} \times \begin{bmatrix} 9.246\ 4 & -5.347\ 1 & -3.899\ 3 \\ -5.347\ 1 & 7.420\ 2 & -2.073\ 1 \\ -3.899\ 3 & -2.073\ 1 & 5.972\ 4 \end{bmatrix},$$

与

$$\hat{\mathbf{P}}/n = 10^{-5} \times \begin{bmatrix} 9.260\ 3 & -5.231\ 7 & -4.028\ 6 \\ -5.231\ 7 & 7.581\ 7 & -2.350\ 0 \\ -4.028\ 6 & -2.350\ 0 & 6.378\ 6 \end{bmatrix}。$$

这两个协方差矩阵估计值 $\mathbf{P}_0/n$ 与$\hat{\mathbf{P}}/n$，可用来计算设计效应矩阵估计值$\hat{\mathbf{D}}$和式 7.12 与式 7.17 中的皮尔逊与内曼检验统计量。注意，在计算式 7.9 中的X^2_{des}，X^2_{P} 与 X^2_{N} 时，我们并不需要完整的矩阵，而可以使用估计值$\hat{\mathbf{V}}_{des}$中的 2 × 2 子矩阵——对应于向量$\hat{\mathbf{p}}$与 $\mathbf{p}_0$ 的两个元素的 $\mathbf{p}_0/n$ 与$\hat{\mathbf{p}}/n$。当然，皮尔逊与内曼统计量的计算，也可以使用等式 7.12 与式 7.17 中给出的标准公式。

对于修正的皮尔逊与内曼检验统计量，我们得到

$$\hat{d}_{\bullet} = \sum_{j=1}^{3} \hat{d}_j/3 = 1.21$$

$$\hat{\delta}_{\bullet} = \sum_{j=1}^{3} \hat{p}_j p_{0j}^{-1}(1 - \hat{p}_j)\hat{d}_j/2 = 1.17$$

$$1 + \hat{a}^2 = 2\,699^2 \sum_{j=1}^{3}\sum_{k=1}^{3} (\hat{v}_{des}^2(\hat{p}_j, \hat{p}_k)/p_{0j}p_{0k})/(2 \times 1.17^2) = 1.37$$

$$\mathrm{df}_S = (u - 1)/(1 + \hat{a}^2) = 1.46。$$

使用这些估计值,我们得到以下数据:

内曼(多项沃尔德)统计量:

$X_N^2 = 9.96$,自由度为 2,p-值为 0.007。

皮尔逊统计量:

$X_P^2 = 10.15$,自由度为 2,p-值为 0.006。

设计效应均值修正的皮尔逊统计量:

$X_P^2(\hat{d}_{\bullet}) = 10.15/1.21 = 8.38$,自由度为 2,$p$-值为 0.015。

一阶拉奥-斯科特修正的皮尔逊统计量:

$X_P^2(\hat{\delta}_{\bullet}) = 10.15/1.17 = 8.66$,自由度为 2,$p$-值为 0.013。

一阶拉奥-斯科特修正的 F-校正:

$FX_P^2(\hat{\delta}_{\bullet}) = 8.66/2 = 4.33$,自由度为 2 与 24,$p$-值为 0.025。

二阶拉奥-斯科特修正的皮尔逊统计量:

$X_P^2(\hat{\delta}_{\bullet}, \hat{a}^2) = 8.66/1.37 = 6.30$,自由度为 $2/1.37 = 1.46$,p-值为 0.023。

基于设计的沃尔德统计量:

$X_{des}^2 = 15.28$,自由度为 2,p-值为 0.001。

F 校正的沃尔德统计量:

$F_{1.des} = (24 - 3 + 2)/(24 \times 2) \times 15.28 = 7.32$,自由度为 2 与 23,$p$-值为 0.003,

$F_{2.des} = 15.28/2 = 7.64$,自由度为 2 与 24,$p$-值为 0.003。

在这些检验统计量中,我们期望,二阶拉奥-斯科特修正与 F-校正的沃尔德统计量给出最满意的检验结果。只有在表 7.1 中仅有设计效应估计值而没有协方差矩阵估计值 $\hat{\mathbf{V}}_{des}$ 时,才使用设计效应均值的修正与一阶拉奥-斯科特的修正。

检验的结果显示与期望一致,相对于修正的皮尔逊统计量而言,修正的皮尔逊与内曼统计量给出了更大的结果。在修正的检验中,二阶拉奥-斯科特修正与一阶拉奥-斯科特修正的 F-校正最为保守。基于设计的沃尔德统计量则是意料外的取值较大,而其 F-校正也没有显著提高。这一结果可能是由于 $\hat{\mathbf{V}}_{des}$ 较小的自由度($f = 24$)引起了不稳定。事实上,相关的 $\hat{\mathbf{V}}_{des}$ 的 2×2 子矩阵的特征值为 0.000 255 2 与 0.000 013 5,条件数为 18.9,并不显示严重的不稳定问题。

在$\hat{\mathbf{V}}_{des}$的自由度较小的 MFH 调查中，应当选择 7 个检验统计量中的哪一个来弥补整群效应呢？首先假定给出了$\hat{\mathbf{V}}_{des}$的估计值，那么应当选择二阶拉奥-斯科特修正。这是因为，$\hat{\mathbf{V}}_{des}$的非正交性以及二阶修正对不稳定性并不敏感。虽然基于设计的沃尔德统计量及其 F-校正也是渐进有效的，但在这里由于它们取值过大，应当排除它们。应当注意到，在其他样本整群数目较大的检验情形下，基于设计的沃尔德统计量是一个合理的替代品。如果没有协方差矩阵估计值$\hat{\mathbf{V}}_{des}$，而是给出了合适的设计效应估计值，则应当选择一阶拉奥-斯科特修正的 F-校正。看起来，它也成功地降低了不稳定的影响。

检验结果并不支持，样本与人口总体的年龄分布相同的结论。对于基于设计的标准化残差$\hat{e}_j$ 的残差分析显示，最大的偏差在第三个年龄组，标准化残差超出了 $N(0, 1)$分布 1% 的临界值 2. 33。除了设计效应估计值明显小于 1 的最后一个年龄组以外，这些残差小于多项统计量的残差。

拒绝 H_0 意味着，更为合理的是对 MFH 数据加权，使得样本年龄分布与总体年龄分布更好的相配。在章节 5. 1 中，我们演示了怎样得出合适的后续分层权重。我们曾指出，相较于未加权，加权估计值中导致一些细小的差异。在回应变量中，这些差异显然是因年龄而引起的。

7. 3 二维表格检验的预备知识

在二维表格中，同质性检验用来检查回应定类变量在各类别的比例，在某一预测定类变量的各个类别中是否相等。而独立性检验用来检验两个回应定类变量间是否有非 0 的联系。这两种检验在概念上假设的构建不同，在检验结果的解释上也不同。在简单随机抽样下，诸如皮尔逊检验的基于多项式的检验，可以在两种假设中使用同一公式。对于涉及到整群的更复杂的设计，我们在技术上剥离这两种检验，并推导出相应检验统计量的不同的修正。我们首先用 MFH 调查中的一个简单例子来介绍这些检验的预备知识。

独立性检验

让我们首先考虑二维表格独立性检验中最简单的例子。从规模为 $n =$ 2 699人的 MFH 调查的演示数据中，我们有以下两个定类变量——PHYS（工作中身体健康危害性，0：无，1：有些）与 SYSBP（收缩血压，≤ 159 或者 > 159）——的频次表：

PHYS	SYSBP		合 计
	≤159	>159	
0	1 857	362	2 219
1	390	90	480
合计	2 247	452	2 699

对于一个独立性假设,我们的问题是,这两个变量是否有联系。这样,得出原假设,

$$H_0: p_{jk} = p_{j+}p_{+k}, \quad j, k = 1, 2,$$

其中,p_{jk}是未知总体方格比例,p_{j+}与p_{+k}是方格频次为N_{jk}的N个元素总体的相应的行与列的边缘比例。因此,我们有,

$$p_{jk} = N_{jk}/N, p_{11} + p_{12} + p_{21} + p_{22} = 1,$$

$$p_{j+} = p_{j1} + p_{j2}, p_{+k} = p_{1k} + p_{2k}。$$

由于方格与边缘比例的限制条件,原假设简化为 $H_0: p_{11} = p_{1+}p_{+1}$,检验的自由度为 1。

对于独立性假设,观测到的方格与边缘比例$\hat{p}_{jk} = \hat{n}_{jk}/n, \hat{p}_{j+} = \hat{p}_{j1} + \hat{p}_{j2}, \hat{p}_{+k} = \hat{p}_{1k} + \hat{p}_{2k}$可以从观测到的方格频次$\hat{n}_{jk}$中推导得出:

PHYS	SYSBP		合 计
	≤159	>159	
0	0. 688 0	0. 134 2	0. 822 2
1	0. 144 5	0. 033 3	0. 1 778
合计	0. 832 5	0. 167 5	1

注意,整个表格的方格比例合计为 1。独立性假设的皮尔逊检验统计量为,

$$X_P^2(I) = n\sum_{j=1}^{2}\sum_{k=1}^{2}\frac{(\hat{p}_{jk} - \hat{p}_{j+}\hat{p}_{+k})^2}{\hat{p}_{j+}\hat{p}_{+k}} = \frac{n(\hat{p}_{11} - \hat{p}_{1+}\hat{p}_{+1})^2}{\hat{p}_{1+}(1 - \hat{p}_{1+})\hat{p}_{+1}(1 - \hat{p}_{+1})},$$

它是在原独立性假设下观测比例与期望比例之差的平方的换算值。对于原假设下的标准推论,皮尔逊统计量满足自由度为 1 的卡方分布。从上表计算得出,$X_P^2(I) = 1.68$,p-值为 0. 195,明显表示接受原独立性假设。

同质性检验

对于独立性假设,事实上分类变量 SYSBP 与 PHYS 都被当成回应变量。我们也可以从另外一个角度来看待频次表。如果我们将 SYSBP 当成回应变量,将 PHYS 当成预测变量,对于一个同质性假设而言,我们的问题是,SYSBP 在 PHYS 的两个类别上的分布是否相同。这样,就得出原假设,对于两个取值

$k=1,2$,有

$$H_0:p_{1k} = p_{2k}$$

与独立性假设相比,这一假设对应着不同的总体比例。我们有,

$$p_{11}+p_{12}=1, p_{21}+p_{22}=1。$$

由于这样的限制条件,原假设简化为,$H_0:p_{11}=p_{21}$,其自由度也为1。

对于同质性假设,其中的行边缘频次为$\hat{n}_1=\hat{n}_{11}+\hat{n}_{12}$与$\hat{n}_2=\hat{n}_{21}+\hat{n}_{22}$,观测到的边缘比例为$\hat{p}_{j+}=1$, $\hat{p}_{+k}=(\hat{n}_{1k}+\hat{n}_{2k})/n$。而观测到的方格的比例$\hat{p}_{1k}=\hat{n}_{1k}/\hat{n}_1$与$\hat{p}_{2k}=\hat{n}_{2k}/\hat{n}_2$ 则如下表:

PHYS	SYSBP		合计
	≤159	>159	
0	0.836 9	0.163 1	1
1	0.812 5	0.187 5	1
合计	0.832 5	0.167 5	1

注意,行边缘比例$\hat{p}_{1+}$与列边缘比例$\hat{p}_{2+}$均为1。同质性假设的皮尔逊检验统计量如下,

$$X_P^2(H) = \sum_{j=1}^{2}\sum_{k=1}^{2}\frac{\hat{n}_j(\hat{p}_{jk}-\hat{p}_{+k})^2}{\hat{p}_{+k}}$$

$$=\frac{(\hat{p}_{11}-\hat{p}_{21})^2}{\hat{p}_{+1}(1-\hat{p}_{+1})/\hat{n}_1+\hat{p}_{+2}(1-\hat{p}_{+2})/\hat{n}_2},$$

同样的,它是在原同质性假设下,观测比例与期望比例之差的平方的一个换算值。对于原假设下的推论,这一皮尔逊统计量也满足自由度为1的卡方分布。虽然,$X_P^2(H)$与$X_P^2(I)$的公式写法不同,但$X_P^2(H)$的观测值1.68与独立性检验中的相同。同时,接受原假设的结论也是正确的。

方格设计效应

以上的独立性与同质性的皮尔逊检验中假定了简单随机样本。当我们考虑到整群效应时,是否仍然是以上的结论?在独立性与同质性下,可以通过计算观测值的方格与边缘比例的估计值的设计效应估计,来检验这一问题的答案。表7.2将有所帮助。DEFF的第一列是独立性假设下的方格设计效应,而DEFF的第二列则是针对同质性假设的。

表 7.2　MFH 调查中独立性与同质性假设的方格与边缘百分比与设计效应

身体健康风险	收缩血压	独立性检验		同质性检验	
		方格百分比	方格百分比 Deff	行百分比	行百分比 Deff
无	≤159	68.8	1.50	83.7	0.88
	>159	13.4	0.81	16.3	0.88
有	≤159	14.5	1.43	81.3	1.15
	>159	3.3	1.34	18.7	1.15

显然，当设计效应的估计值平均大于 1 时，则需要比未修正的统计量更为保守的、修正过的检验统计量。因此，仍然是接受原假设的结论。独立性假设的方格设计效应估计值的平均值为$\hat{d}_{\bullet}(I)=1.27$，以此得出的修正后的皮尔逊统计量为 $X_P^2(I,\ \hat{d}_{\bullet})=1.32$，其 p-值为 0.251。同质性假设的方格设计效应估计值的平均值为$\hat{d}_{\bullet}(H)=1.01$，以此得出的修正后的皮尔逊统计量为 $X_P^2(H,\ \hat{d}_{\bullet})=1.66$，其 p-值为 0.198。这些基于设计的检验并不产生新的推论结论。但是，更重要的是，它们显示了，虽然未修正的统计量数值上相同，由于不同的修正，校正设计效应的修正的皮尔逊检验统计量给出了数值上并不一致的结果。拉奥-斯科特修正的 $X_P^2(I)$ 与 $X_P^2(H)$ 也有差异；独立性与同质性假设的基于设计的沃尔德统计量也不尽相同。

这些检验也表明，在 MFH 调查中，群内相关对独立性检验的影响大于对同质性检验的影响。这可能是因为我们的对象是类别交叉的子群，部分也是因为方差估计中较小的自由度。应当注意到，也有与此相反的情形。有例子显示，整群的膨胀效应在某些调查中，对独立性检验的影响小于对同质性检验的影响(Rao and Thomas，1988)。特别是在预测变量的类别是互不交叉型的区域时，更是如此。

与简单拟合度检验一样，对于复杂调查中更普遍的 $r\times c$ 表格，可以构建基于设计的沃尔德统计量的 F-校正和标准皮尔逊与内曼检验统计量的二阶拉奥-斯科特修正来检验独立性与同质性。在对已公布表格的二手分析中，如果仅给出了方格与边缘设计估计值，而没有给出比例估计值的基于设计的协方差矩阵估计，则可以使用设计效应均值修正与一阶拉奥-斯科特修正。

7.4　同质性检验

在调查分析文献中，同质性检验通常用来检验一个回应变量在一组互不

交叉的地域上的分布的同质性。这样的样本使用了多级抽样设计独立抽取(例如,Rao and Thomas,1988)。这些地域假定是相互区隔的,同一样本整群的所有元素均在同一个地域内(预测变量的类别)。回应变量的类别通常是交叉横切地域的。更普遍的,同质性检验可以被看成是最简单的对数模型。其回应变量是二分或多值变量,预测变量在实际中并不局限于仅仅是相互区隔的类别变量。

在同质性检验中,假定表格的列由回应变量的类别组成,行由地域组成,并且,假定每一行的方格比例合计为1。这样,总体表格如下:

地域	回应变量						合计
	1	2	…	k	…	c	
1	p_{11}	p_{12}	…	p_{1k}	…	p_{1c}	1
2	p_{21}	p_{22}	…	p_{2k}	…	p_{2c}	1
⋮	⋮	⋮		⋮		⋮	⋮
j	p_{j1}	p_{j2}	…	p_{jk}	…	p_{jc}	1
⋮	⋮	⋮		⋮		⋮	⋮
r	p_{r1}	p_{r2}	…	p_{rk}	…	p_{rc}	1

为了简便,我们仅考虑两个地域的例子,并假定地域是相互区隔的类别。一个 c 个类别的回应变量的 $r=2$ 地域的同质性假设,在章节 7.3 中已经给出为,$H_0:p_{1k}=p_{2k}$,其中,$p_{1k}=N_{1k}/N_1$ 与 $p_{2k}=N_{2k}/N_2$ 是第一地域与第二地域的未知总体比例,$k=1,\cdots,c$。使用向量,这一假设可以写成,$H_0:\mathbf{p}_1=\mathbf{p}_2$,其中,$\mathbf{p}_j=(p_{j1},\cdots,p_{j,c-1})'$表示地域$j$中行比例$p_{jk}$的总体向量。由于每一地域的比例必须独立地合计为1,所以每一个地域比例向量有 $c-1$ 个元素。另外,我们用 $\mathbf{p}=(p_{+1},\cdots,p_{+,c-1})'$表示在 H_0 下的未知普通比例向量,其中,$p_{+k}=N_{+k}/N$,$N_{+k}=N_{1k}+N_{2k}$。

根据从地域中得到的独立样本,估计的地域比例向量用$\hat{\mathbf{p}}_j=(\hat{p}_{j1},\cdots,\hat{p}_{j,c-1})'$表示。其中,$\hat{p}_{jk}=\hat{n}_{jk}/\hat{n}_j$是相应总体比例$p_{jk}$的一个一致的估计值,$\hat{n}_{jk}$与$\hat{n}_j$是调整了选中概率不等与无应答情形的加权换算过的边缘频次,而$\sum_{k=1}^{c}\hat{n}_{jk}=\hat{n}_j$。与拟合度检验中一样,当我们的对象——地域样本的子集的规模事先并不固定时,$\hat{p}_{jk}$是一个比率估计值。对于从 MFH 与 OHC 调查中得来的演示数据,也是如此。

基于设计的沃尔德统计量

让我们用$\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)$来表示第一个地域的比例估计向量$\hat{\mathbf{p}}_1$的协方差矩阵的

一致估计值，$\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2)$来表示第二个地域的相同估计值。各个地域的协方差矩阵的估计值方法与拟合度中的方法相似。使用$\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)$与$\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2)$，由于类别间互不交叉并且 $r=2$，两个地域的同质性假设的基于设计的沃尔德统计量 X^2_{des} 如下，

$$X^2_{des} = (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2)'(\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1) + \hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2))^{-1}(\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2), \qquad (7.20)$$

这一沃尔德统计量符合自由度为$(2-1)\times(c-1)=(c-1)$的渐进卡方分布。同时，当 $c=2$ 时，X^2_{des}简化为 $X^2_{des}=(\hat{p}_{11}-\hat{p}_{21})^2/(\hat{v}_{des}(\hat{p}_{11})+\hat{v}_{des}(\hat{p}_{21}))$。式 7.20 中的 X^2_{des}并不能直接扩展到多于两个地域的情形，那要复杂得多(Rao and Thomas，1988)。

当各个地域内的整群样本数目较大时，统计量 X^2_{des}相对较为合理。但情况不是如此时，将遇到不稳定的问题。这时，可以使用 F-校正的沃尔德统计量。使用$f=m-H$ 作为估计值$(\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)+\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2))$整体上的自由度——$m$ 与 H 是两个地域中的总样本整群数与层级数，这一校正值为，

$$F_{1.des} = \frac{f-(c-1)+1}{f(c-1)}X^2_{des}, \qquad (7.21)$$

它满足自由度为$(c-1)$与$(f-(c-1)+1)$的 F-分布。同时，更进一步，

$$F_{2.des} = X^2_{des}/(c-1), \qquad (7.22)$$

它满足自由度为$(c-1)$与f的 F-分布。当与回应变量的类别数目 c 相比，f 并不太大时，这些检验统计量可以有效地降低不稳定性的影响。

对皮尔逊与内曼检验统计量的修正

在 $r=2$ 个地域情形下，同质性假设的皮尔逊检验统计量是，

$$X^2_{P} = \sum_{j=1}^{2}\sum_{k=1}^{c}\frac{\hat{n}_j(\hat{p}_{jk}-\hat{p}_{+k})^2}{\hat{p}_{+k}}$$

$$= (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2)'(\hat{\mathbf{P}}/\hat{n}_1 + \hat{\mathbf{P}}/\hat{n}_2)^{-1}(\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2), \qquad (7.23)$$

其中，$\hat{p}_{+k}=(\hat{n}_1\hat{p}_{1k}+\hat{n}_2\hat{p}_{2k})/(\hat{n}_1+\hat{n}_2)$是表格各列的边缘比例估计值。即，$H_0$ 下的假定相同比例向量 $\mathbf{p}$ 的元素 p_{+k}的估计值，并且，$\hat{\mathbf{P}}=\mathrm{diag}(\hat{\mathbf{p}})-\hat{\mathbf{p}}\hat{\mathbf{p}}'$，$\hat{\mathbf{P}}/\hat{n}_1$是第一个地域估计向量$\hat{\mathbf{p}}$的多项协方差矩阵估计值，$\hat{\mathbf{P}}/\hat{n}_2$ 则对应着第二个地域。同时，当 $c=2$ 时，X^2_P 简化为$\hat{n}_1\hat{n}_2(\hat{p}_{11}-\hat{p}_{21})^2/((\hat{n}_1+\hat{n}_2)\hat{p}_{+1}(1-\hat{p}_{+1}))$。

也可以取而使用内曼检验统计量。它可以通过假定在两个地域内独立的多项抽样，从基于设计的沃尔德统计量(式 7.20)中推导出来：

$$X^2_{N} = \sum_{j=1}^{2}\sum_{k=1}^{c}\frac{\hat{n}_j(\hat{p}_{jk}-\hat{p}_{+k})^2}{\hat{p}_{jk}}$$

$$= (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2)'(\hat{\mathbf{P}}_1/\hat{n}_1 + \hat{\mathbf{P}}_2/\hat{n}_2)^{-1}(\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2), \qquad (7.24)$$

其中，$\hat{\mathbf{P}}_1 = \text{diag}(\hat{\mathbf{p}}_1) - \hat{\mathbf{p}}_1\hat{\mathbf{p}}_1'$，$\hat{\mathbf{P}}_1/\hat{n}_1$ 是第一个地域的多项协方差矩阵估计值，$\hat{\mathbf{P}}_2/\hat{n}_2$ 则对应着第二个地域。同时，当 $c=2$ 时，X_N^2 简化为$(\hat{p}_{11}-\hat{p}_{21})^2/[\hat{p}_{11}(1-\hat{p}_{11})/\hat{n}_1+\hat{p}_{21}(1-\hat{p}_{21})/\hat{n}_2]$。注意，$X_P^2$ 和 X_N^2 的公式与基于设计的沃尔德统计量的公式很相似，仅有的区别在于使用了哪一个协方差矩阵估计值。

皮尔逊与内曼检验统计量在简单随机抽样中是有效的，它们满足自由度为$(c-1)$的卡方分布。但在更复杂的设计中，它们需要校正整群效应的调整。调整与拟合度检验中相似，但从技术上而言，使用的公式不一样。

对于 X_P^2 和 X_N^2 的设计效应均值修正与一阶拉奥-斯科特修正，需要两个地域的方格设计效应估计值；对于二阶拉奥-斯科特修正，则需要通用的设计效应矩阵估计值。地域 j 的设计效应估计值的形式为，

$$\hat{d}_{jk} = \hat{d}(\hat{p}_{jk}) = \hat{n}_j\hat{v}_{jk}/(\hat{p}_{+k}(1-\hat{p}_{+k})),j=1,2,k=1,\cdots,c,$$

其中，$\hat{v}_{1k}$是第一个地域的协方差矩阵估计值$\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)$的第 k 个对角元素，$\hat{v}_{2k}$是对应于第二个地域中$\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2)$的元素。通用的设计效应矩阵估计值为，

$$\hat{\mathbf{D}} = \frac{\hat{n}_1\hat{n}_2}{\hat{n}_1+\hat{n}_2}\hat{\mathbf{P}}^{-1}(\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)+\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2))。\qquad(7.25)$$

皮尔逊与内曼检验统计量的设计效应均值修正值为：

$$X_P^2(\hat{d}_\bullet) = X_P^2/\hat{d}_\bullet,X_N^2(\hat{d}_\bullet) = X_N^2/\hat{d}_\bullet,\qquad(7.26)$$

其中，

$$\hat{d}_\bullet = \sum_{j=1}^{2}\sum_{k=1}^{c}\hat{d}_{jk}/(2c)$$

是设计效应估计值的均值。使用$\hat{\mathbf{D}}$中的特征值$\hat{\delta}_k$，式 7.23 与式 7.24 中的皮尔逊与内曼检验统计量的一阶拉奥-斯科特修正给出如下：

$$X_P^2(\hat{\delta}_\bullet) = X_P^2/\hat{\delta}_\bullet,X_N^2(\hat{\delta}_\bullet) = X_N^2/\hat{\delta}_\bullet,\qquad(7.27)$$

其中，

$$\hat{\delta}_\bullet = \text{tr}(\hat{\mathbf{D}})/(c-1) = \frac{1}{c-1}\sum_{j=1}^{2}\left(1-\frac{\hat{n}_j}{\hat{n}_1+\hat{n}_2}\right)\sum_{k=1}^{c}\frac{\hat{p}_{jk}}{\hat{p}_{+k}}(1-\hat{p}_{jk})\hat{d}_{jk}$$

是未知通用设计效应矩阵 $\mathbf{D}$ 的特征值 δ_k 的均值δ的估计值。注意，估计值$\hat{\delta}_\bullet$也可以通过首先计算对角元素的和，即是迹，直接从$\hat{\mathbf{D}}$中获得。当设计效应估计值或是特征值变化并不剧烈时，这两种修正较为合理。从这个意义上讲，它们是相近的。

当特征值估计$\hat{\delta}_k$变动较大时，X_P^2 和 X_N^2 的二阶拉奥-斯科特修正更优。对于皮尔逊统计量，这一修正给出为，

$$X_P^2(\hat{\delta}_\bullet,\hat{a}^2) = X_P^2(\hat{\delta}_\bullet)/(1+\hat{a}^2),\qquad(7.28)$$

其中，$\hat{a}^2$ 是特征值估计$\hat{\delta}_k$ 的离异系数的平方，它从以下公式得到，

$$\hat{a}^2 = \sum_{k=1}^{c-1} \hat{\delta}_k^2 / ((c-1)\hat{\delta}_{\bullet}^2) - 1,$$

这里的特征值平方和可以当成二阶的通用设计效应矩阵估计值的迹：

$$\sum_{k=1}^{c-1} \hat{\delta}_k^2 = \mathrm{tr}(\hat{\mathbf{D}}^2)。$$

皮尔逊检验统计量的二阶拉奥-斯科特修正满足自由度为萨特斯韦特校正 $\mathrm{df}_S = (c-1)/(1+\hat{a}^2)$的渐进卡方分布。对式 7.27 中内曼统计量的一阶修正 $X_N^2(\hat{\delta}_{\bullet})$也可以做相似的调整。

当地域协方差矩阵估计值$\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)$与$\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2)$是基于相对较小数目的样本整群时，它们可能并不稳定。因而，可以使用一阶检验统计量的 F-校正。式 7.27 中两个地域的皮尔逊统计量的 F-校正为，

$$FX_P^2(\hat{\delta}_{\bullet}) = X_P^2(\hat{\delta}_{\bullet})/(c-1) \tag{7.29}$$

它是满足自由度为$(c-1)$与f 的 F-分布。内曼统计量的纠正与此相似。

残差分析

当拒绝原同质性假设时，可以计算标准化残差来检查方格与假设的比例偏离有多远。使用方格设计效应估计值$\hat{d}_{jk}$，我们计算基于设计的标准化残差，

$$\hat{e}_{jk} = (\hat{p}_{jk} - \hat{p}_{+k})/\mathrm{s.e}_{des}(\hat{p}_{jk} - \hat{p}_{+k}), \quad j = 1,2, k = 1,\cdots,c, \tag{7.30}$$

其中，原始残差的标准误估计值 $\mathrm{s.e}_{des}(\hat{p}_{jk} - \hat{p}_{+k})$，从基于设计的方差估计值获得。第一个地域的方差估计值为，

$$\hat{v}_{des}(\hat{p}_{1k} - \hat{p}_{+k}) = \frac{\hat{n}_2(\hat{n}_2\hat{d}_{1k} + \hat{n}_1\hat{d}_{2k})}{(\hat{n}_1 + \hat{n}_2)^2}\hat{p}_{+k}(1 - \hat{p}_{+k})/\hat{n}_1, \quad k = 1,\cdots,c,$$

第二个地域的方差估计值为，

$$\hat{v}_{des}(\hat{p}_{2k} - \hat{p}_{+k}) = \frac{\hat{n}_1(\hat{n}_2\hat{d}_{1k} + \hat{n}_1\hat{d}_{2k})}{(\hat{n}_1 + \hat{n}_2)^2}\hat{p}_{+k}(1 - \hat{p}_{+k})/\hat{n}_2, \quad k = 1,\cdots,c,$$

注意，在简单随机抽样下，$\hat{d}_{1k} = \hat{d}_{2k} = 1$，第一个地域的方差估计值简化为，

$$\hat{v}_{srs}(\hat{p}_{1k} - \hat{p}_{+k}) = \frac{\hat{n}_2}{\hat{n}_1 + \hat{n}_2}\hat{p}_{+k}(1 - \hat{p}_{+k})/\hat{n}_1, \quad k = 1,\cdots,c,$$

第二个地域的方差估计值简化为，

$$\hat{v}_{srs}(\hat{p}_{2k} - \hat{p}_{+k}) = \frac{\hat{n}_1}{\hat{n}_1 + \hat{n}_2}\hat{p}_{+k}(1 - \hat{p}_{+k})/\hat{n}_2, \quad k = 1,\cdots,c,$$

从这些公式中可以推出，在正的群内相关的情形下，得到的基于设计的标准化残差要小于在简单随机抽样假设下的标准化残差。基于设计的标准化残差可

以用来比照标准 $N(0, 1)$ 分布的临界值。

范例 7.2

OHC 调查中两个总体的同质性检验。我们考虑检验变量 PSYCH 的类别比例的同质性。它是测量整体精神压力的 9 个心理症状的第一个主成分,并被差不多等分成 3 个类别。两个总体的区分是单位类型,公共服务部门组成第一个子群,所有其余公司则划入第二个子群(表 7.3)。注意,分组是依据工业上的分层方法,并且是互不交叉的。所以,在各个总体中的样本可以假定为独立抽取的。在 250 个样本整群中,49 个来自第一子群,201 个来自第二子群。两个子群中的元素数据均是自加权的。

表 7.3 OHC 调查中公共服务部门与其他部门心理症状组别比例(括号内为设计效应)

部门类型	PSYCH			合计	样本规模
	1	2	3		
公共服务	0.293 9 (2.02)	0.334 5 (1.24)	0.371 6 (1.74)	1.00	1 184
其他部门	0.352 6 (1.73)	0.321 6 (1.23)	0.325 8 (1.57)	1.00	6 657
所有部门	0.343 7	0.323 6	0.332 7	1.00	7 841

在公共服务部门中,得到了比其他部门更高比例的严重心理症状(第 3 类别)。为了检验这一变化趋势,两个总体的类别比例的同质性假设为 $H_0: p_{1k} = p_{2k}, k = 1, 2, 3$。方格设计效应的均值为 1.59,显示中等程度的整群效应。在检验过程中应做相应的调整。为了计算有效的检验统计量,我们首先得到两个完整的协方差矩阵估计值 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)$ 与 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2)$。它们是,

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1) = 10^{-5} \times \begin{bmatrix} 35.339\,4 & -12.140\,8 & -23.198\,6 \\ -12.140\,8 & 23.357\,0 & -11.216\,1 \\ 23.198\,6 & -11.216\,1 & 34.414\,8 \end{bmatrix},$$

与

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2) = 10^{-5} \times \begin{bmatrix} 5.917\,7 & -2.397\,8 & -3.520\,0 \\ -2.397\,8 & 4.041\,7 & -1.643\,9 \\ -3.520\,0 & -1.643\,9 & 5.163\,9 \end{bmatrix}。$$

由于 $c - 1 = 2$,我们在计算沃尔德统计量与拉奥-斯科特修正值时,使用 PSYCH 的头两个类别以及估计值 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)$ 与 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2)$ 的第一个 2 × 2 矩阵。对于同质性的基于设计的沃尔德统计量,我们得到 $X^2_{des} = 8.62$,其自由度为 2,p-值为 0.013 4。这表示各个总体间的比例并不相等。式 7.21 与式 7.22 中的

X^2_{des}的 F-校正给出，$F_{1.des}=4.29$，满足自由度为 2 与 244 的 F-分布，p-值为 0.014 7；$F_{2.des}=4.31$，满足自由度为 2 与 245 的 F-分布，p-值为 0.014 4。这两个校正对于 X^2_{des}没有太大的影响。这是因为样本整群数目相对较大，$\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)$ 与$\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2)$可以假定比较稳定。

我们计算式 7.23 与式 7.24 中皮尔逊与内曼检验统计量的二阶拉奥-斯科特修正，并作为另一种有效的检验方法。未修正的统计量得出 $X^2_P=16.93$，p-值为 0.000 2；$X^2_N=17.77$，p-值为 0.000 1，两者均在 0.001 水平显著。与期望的相符，相对于 X^2_{des}，它们较大。对于拉奥-斯科特修正，首先得到式 7.25 中的通用的设计效应矩阵估计值：

$$\hat{\mathbf{D}}=\begin{bmatrix}2.013\,74 & -0.036\,63\\0.355\,54 & 1.239\,77\end{bmatrix}。$$

$\hat{\mathbf{D}}$ 的对角元素的均值$\hat{\delta}_{\bullet}=\mathrm{tr}(\hat{\mathbf{D}})/2=1.627$，特征值的平方和为 $\sum_{k=1}^{2}\hat{\delta}_k^2=\mathrm{tr}(\hat{\mathbf{D}}^2)=5.566$。因此，二阶修正因子为$(1+\hat{a}^2)=1.052$，其萨特斯韦特调整的自由度 $\mathrm{df}_S=1.902$。由此得出，$X^2_P(\hat{\delta}_{\bullet},\hat{a}^2)=9.89$，$p$- 值为 0.006 3，以及 $X^2_N(\hat{\delta}_{\bullet},\hat{a}^2)=10.38$，$p$- 值为 0.004 9。两者均在 0.01 水平显著。这样的结果比沃尔德检验的结果更大。这也显示，与 MFH 的例子（见范例7.1）不同，基于设计的沃尔德在 OHC 例子中的结果较为满意。

在仅有如表 7.3 中的信息来检验同质性的前提下，我们最后来计算 X^2_P 与 X^2_N 在式 7.26 与式 7.27 中的一阶修正。设计效应的均值估计值为$\hat{d}_{\bullet}=1.59$，相应的 X^2_P 与 X^2_N 的修正为 $X^2_P(\hat{d}_{\bullet})=10.66$，$p$-值为 0.004 8，以及 $X^2_N(\hat{d}_{\bullet})=11.19$，$p$-值为 0.003 7。两者均在 0.01 水平显著。使用方格设计效应估计值以及方格比例，我们得到$\hat{\delta}_{\bullet}=1.627$。一阶拉奥-斯科特修正为 $X^2_P(\hat{\delta}_{\bullet})=10.41$，$p$-值为 0.005 5，以及 $X^2_N(\hat{\delta}_{\bullet})=10.92$，$p$-值为 0.004 3。两者也均在 0.01水平显著。根据 X^2_P 与 X^2_N 在式 7.29 中的 F-校正得出 $FX^2_P=5.20$，p-值为 0.006 1，以及 $FX^2_N=5.46$，p-值为 0.004 8，显示与其他一阶修正的结果没有明显的变化。这也再一次显示了检验结果的稳定性。

由于所有的检验至少在 0.05 水平上拒绝了 H_0，我们计算两个类别的基于设计的标准化残差$\hat{e}_{jk}$。使用公式 7.30，它们如下：

PSYCH	公共服务$\hat{e}_{1k}$	其他部门$\hat{e}_{2k}$
1	-2.79	2.79
2	0.78	-0.78
3	2.35	-2.35

公共服务部门与其他公司的残差总和为0。注意,从标准化残差的绝对值看,最大的在PSYCH的第1与第3类别。在PSYCH的第3类别,差异的方向有利于公共服务部门;而在第1类别,则相反。基于设计的标准化残差,也超过了这些类别中标准正态分布$N(0, 1)$1%的临界值2.33。

根据所有相关信息,我们得出结论,基于设计的沃尔德统计量能够为同质性假设提供满意与可用的检验方法。当给出方格设计效应,但没有两个地域的协方差矩阵估计值时,我们可以选择皮尔逊或是内曼检验统计量的拉奥-斯科特修正。但在考虑的例子中,推论的结论与所选的检验统计量没有关系;而拒绝两个总体中PSYCH比例同质的原假设的结论的程度,却大小有所不同。

对数模型为同质性假设检验提供了一个方便通用的框架。在2×3表格中,INDU(公司类型)类别间PSYCH比例同质假设的检验,可以被当成一个简单的多值回应变量的对数模型。通过对PSYCH比率对数(logits)拟合完整的对数模型(截距+INDU),并用沃尔德统计量检验INDU系数的显著性,可以得到一个同质性检验。观测到的沃尔德检验统计量为$X^2_{des} = 8.13$,p-值为0.017 1。这一结果虽然略微保守,与前面的沃尔德检验统计量差得不多。

多于两个地域的情形

我们讨论了两个地域的同质性检验,其中的地域由相互区隔的类型组成。对于多于两个地域的情形,可以直截了当地推导出皮尔逊与内曼检验统计量的基于设计的沃尔德统计量与拉奥-斯科特修正。这些推导需要涉及更多的矩阵代数,我们跳过这一推导,让读者参阅拉奥与托马斯(Rao and Thomas,1988)。

相互区隔的类别的同质性检验,是更综合性的任何定类预测变量的检验的特例。含有二分回应变量的这样的情形,在第8章中将用于对数建模讨论。在那里,将放开相互区隔的假设,并将讨论预测变量的类别交叉。这样以来,基于设计的回应变量比例的协方差矩阵,无法像相互隔离的地域例子中,在各个预测变量子群中分开估算。同时,也必须估算地域间的协方差。这一协方差在相互隔离的地域例子中被假定为0。

7.5 独立性检验

独立性检验,用于检查在同一总体中两个定类变量是否存在非0的联系。假定数据从一个边缘总和并不固定的简单总体中抽取,并组成了一个$r \times c$的互联表。因此,假定所有总体比例p_{jk}之和为1。总体表格为,

第一个变量	第二个变量						合计
	1	2	…	k	…	c	
1	p_{11}	p_{12}	…	p_{1k}	…	p_{1c}	p_{1+}
2	p_{21}	p_{22}	…	p_{2k}	…	p_{2c}	p_{2+}
⋮	⋮	⋮		⋮		⋮	⋮
j	p_{j1}	p_{j2}	…	p_{jk}	…	p_{jc}	p_{j+}
⋮	⋮	⋮		⋮		⋮	⋮
r	p_{r1}	p_{r2}	…	p_{rk}	…	p_{rc}	p_{r+}
合计	p_{+1}	p_{+2}	…	p_{+k}	…	p_{+c}	1

在提出原假设与解释检验结果时,注意到以下方面很重要:我们研究的是一个对称的情形,没有变量被假定为预测变量。有着 r 与 c 个类别的两个回应变量通常是类别交叉或是类别混杂的。这样,它们才能横切层级与整群。在章节 7.3 中给出了回应变量的独立性假设,$H_0:p_{jk} = p_{j+}p_{+k}$。其中,$p_{jk} = N_{jk}/N$,$p_{j+} = \sum_{k=1}^{c} p_{jk}$ 与 $p_{+k} = \sum_{j=1}^{r} p_{jk}$ 是边缘比例,$j = 1, \cdots, r, k = 1, \cdots, c$。显然,在原假设下,当实际的未知方格比例 p_{jk} 与期望的方格比例 $p_{j+}p_{+k}$ 很接近时,这两个变量间相互独立。在构造合适的独立性假设检验统计量时,使用了这一事实。

为了推导出独立性的检验统计量,我们将原假设写成等同的格式,$H_0:F_{jk} = p_{jk} - p_{j+}p_{+k} = 0$。其中,$j = 1, \cdots, r-1, k = 1, \cdots, c-1$。这是因为限制条件 $\sum_{j=1}^{r}\sum_{k=1}^{c} p_{jk} = 1$。因而,$F_{jk}$ 是原假设下未知方格比例与期望比例的差值残差,而原假设则是差值残差为 0。残差可以组成一个含有 $(r-1)\times(c-1)$ 行的列向量 $\mathbf{F} = (F_{11}, \cdots, F_{1,c-1}, \cdots, F_{r-1,1}, \cdots, F_{r-1,c-1})'$。

从 n 个元素的样本得到的方格比例的估计值 $\hat{p}_{jk} = \hat{n}_{jk}/n$ 是相应的未知比例 p_{jk} 的一致估计值。$\hat{n}_{jk}$ 是考虑个体元素选中概率不等并针对无应答修正的、换算加权过的方格频次,而 $\sum_{j=1}^{r}\sum_{k=1}^{c}\hat{n}_{jk} = n$。当子集的整体样本规模并非事先固定时,$\hat{p}_{jk}$ 是比率估计值。这与 MFH 与 OHC 调查中的演示数据相同。在拟合度与同质性假设中,我们也在此有相同的假设。

协方差矩阵估计值

让我们首先推导在各种抽样设计假设情形下,残差差异的估算向量 $\hat{\mathbf{F}}$ 的协方差矩阵估计值。它将用于获得皮尔逊与内曼检验统计量以及基于设计的沃尔德统计量。残差差异的估算向量为,

$$\hat{\mathbf{F}} = (\hat{F}_{11}, \cdots, \hat{F}_{1,c-1}, \cdots, \hat{F}_{r-1,1}, \cdots, \hat{F}_{r-1,c-1})', \qquad (7.31)$$

其中，$\hat{F}_{jk}=\hat{p}_{jk}-\hat{p}_{j+}\hat{p}_{+k}$，$\hat{p}_{j+}$与$\hat{p}_{+k}$是相应的边缘比例估计值。对于基于设计的沃尔德统计量，我们推导出考虑了抽样设计复杂性的$\hat{\mathbf{F}}$的一致的协方差矩阵估计值$\hat{\mathbf{V}}_F$，其公式如下，

$$\hat{\mathbf{V}}_F = \hat{\mathbf{H}}' \hat{\mathbf{V}}_{des} \hat{\mathbf{H}}, \tag{7.32}$$

其中，$(r-1)(c-1)\times(r-1)(c-1)$矩阵$\hat{\mathbf{H}}$是针对$p_{jk}$求值于$\hat{p}_{jk}$的 $\mathbf{F}$ 的偏微分矩阵。矩阵$\hat{\mathbf{V}}_{des}$是方格比例估计值向量$\hat{\mathbf{p}}=(\hat{p}_{11},\cdots,\hat{p}_{1,c-1},\cdots,\hat{p}_{r-1,1},\cdots,\hat{p}_{r-1,c-1})'$的渐进协方差矩阵 $\mathbf{V}/n$ 的估计值。$\hat{\mathbf{V}}_{des}$的估计值通过前面用于拟合度与同质性假设中的线性化方法获得。在实际中，可以从元素数据中拟合一个以定类变量为模型项的、不含截距的完整交互作用线性模型来计算$\hat{\mathbf{V}}_{des}$。估算的模型系数与观测到的比例相等，系数的协方差矩阵估计值即是$\hat{\mathbf{V}}_{des}$的估计值。

$\hat{\mathbf{F}}$的两个多项协方差矩阵估计值如下。对于皮尔逊检验统计量，我们在原假设下，推导出$\hat{\mathbf{F}}$的期望多项协方差矩阵估计值$\hat{\mathbf{P}}_{0F}/n$，并有

$$\hat{\mathbf{P}}_{0F} = \hat{\mathbf{H}}' \hat{\mathbf{P}}_0 \hat{\mathbf{H}}, \tag{7.33}$$

其中，$\hat{\mathbf{P}}_0=\mathrm{diag}(\hat{\mathbf{p}}_0)-\hat{\mathbf{p}}_0\hat{\mathbf{p}}'_0$，$\hat{\mathbf{p}}_0$ 为原假设下期望比例向量，即元素为$\hat{p}_{j+}\hat{p}_{+k}$的向量。对于内曼检验统计量，我们推导出$\hat{\mathbf{F}}$的观测到的多项协方差矩阵估计值$\hat{\mathbf{P}}_F/n$，

$$\hat{\mathbf{P}}_F = \hat{\mathbf{H}}' \hat{\mathbf{P}} \hat{\mathbf{H}}, \tag{7.34}$$

其中，$\hat{\mathbf{P}}=\mathrm{diag}(\hat{\mathbf{p}})-\hat{\mathbf{p}}\hat{\mathbf{p}}'$。注意，$\hat{\mathbf{F}}$的所有协方差矩阵估计值形式相似，都使用偏微分矩阵$\hat{\mathbf{H}}$。

基于设计的沃尔德统计量

使用残差差异的估计向量$\hat{\mathbf{F}}$，以及从式 7.32 得到的一致的协方差矩阵估计值$\hat{\mathbf{V}}_F$，我们得到独立性假设的基于设计的沃尔德统计量

$$X^2_{des} = \hat{\mathbf{F}}' \hat{\mathbf{V}}_F^{-1} \hat{\mathbf{F}}, \tag{7.35}$$

它满足自由度为$(r-1)(c-1)$的渐进卡方分布。与拟合度和同质性中的沃尔德检验相同，当估计值$\hat{\mathbf{V}}_F$ 仅有较小的自由度f时，这一检验统计量会有不稳定的问题。这时，可以使用 X^2_{des}的 F-校正，其中，

$$F_{1.des} = \frac{f-(r-1)(c-1)-1}{f(r-1)(c-1)}X^2_{des}, \tag{7.36}$$

它满足自由度为$(r-1)(c-1)$与$(f-(r-1)(c-1)-1)$的 F-分布，以及，

$$F_{2.des} = \frac{X^2_{des}}{(r-1)(c-1)}, \tag{7.37}$$

它满足自由度为$(r-1)(c-1)$与f的 F-分布。

皮尔逊与内曼检验统计量的修正

章节 7.3 中独立性假设的皮尔逊检验统计量为，

$$X^2_{\text{P}} = n\sum_{j=1}^{r}\sum_{k=1}^{c}\frac{(\hat{p}_{jk}-\hat{p}_{j+}\hat{p}_{+k})^2}{\hat{p}_{j+}\hat{p}_{+k}}。 \tag{7.38}$$

内曼检验统计量可以用作替代，

$$X^2_{\text{N}} = n\sum_{j=1}^{r}\sum_{k=1}^{c}\frac{(\hat{p}_{jk}-\hat{p}_{j+}\hat{p}_{+k})^2}{\hat{p}_{jk}}。 \tag{7.39}$$

这些统计量的观测值可以从估计的方格与边缘比例中获得。在简单随机抽样下，两个统计量均满足自由度为$(r-1)(c-1)$的渐进卡方分布。

为了构造一个通用的框架，使用相应的矩阵公式，我们将式 7.38 中的皮尔逊检验统计量写成，

$$X^2_{\text{P}} = n\,\hat{\mathbf{F}}'\,\hat{\mathbf{P}}^{-1}_{0F}\,\hat{\mathbf{F}} \tag{7.40}$$

其中使用了式 7.33 中的原始多项协方差矩阵估计值$\hat{\mathbf{P}}_{0F}/n$。而将式 7.39 中的内曼检验统计量写成

$$X^2_{\text{N}} = n\,\hat{\mathbf{F}}'\,\hat{\mathbf{P}}^{-1}_{F}\,\hat{\mathbf{F}} \tag{7.41}$$

其中使用了式 7.34 中的实际多项协方差矩阵估计值$\hat{\mathbf{P}}_F/n$。注意，这两个统计量与式 7.35 中的基于设计的沃尔德统计量 X^2_{des}非常相似，仅有的差别是使用了不同的残差差异的协方差矩阵估计值。应当注意到，在计算 X^2_{des}，X^2_{P} 以及 X^2_{N} 时，向量$\hat{\mathbf{F}}$是一个$(r-1)(c-1)$的列向量，而协方差矩阵则是$(r-1)(c-1)\times(r-1)(c-1)$矩阵。因此，在 2×2 的表格例子中，$\hat{\mathbf{F}}$以及协方差矩阵估计值$\hat{\mathbf{P}}_{0F}$与$\hat{\mathbf{P}}_F$简化成换算因子。

在复杂调查中，与相应的拟合度与同质性检验一样，我们有相似的理由对统计量 X^2_{P} 与 X^2_{N} 做整群效应修正。使用二阶拉奥-斯科特修正，得到的渐进有效的式 7.40 中皮尔逊统计量的修正检验统计量如下，

$$X^2_{\text{P}}(\hat{\delta}_{\bullet},\hat{a}^2) = X^2_{\text{P}}/[\hat{\delta}_{\bullet}(1+\hat{a}^2)] \tag{7.42}$$

其中，

$$\hat{\delta}_{\bullet} = \text{tr}(\hat{\mathbf{D}})/[(r-1)(c-1)]$$

是通用设计效应矩阵估计值的特征值$\hat{\delta}_l$ 的均值。通用设计效应矩阵估计

值为，

$$\hat{\mathbf{D}} = n\,\hat{\mathbf{P}}_{0F}^{-1}\,\hat{\mathbf{V}}_F, \tag{7.43}$$

而

$$\hat{a}^2 = \sum_{l=1}^{(r-1)(c-1)} \hat{\delta}_l^2 / [(r-1)(c-1)\,\hat{\delta}_\bullet^2] - 1$$

是特征值估计值$\hat{\delta}_l$ 的离异系数的平方。特征值的平方和为，

$$\sum_{l=1}^{(r-1)(c-1)} \hat{\delta}_l^2 = \mathrm{tr}(\hat{\mathbf{D}}^2)$$

式 7.42 中的二阶修正统计量满足萨特斯韦特自由度为

$$\mathrm{df}_S = \frac{(r-1)(c-1)}{(1+\hat{a}^2)}$$

的渐进卡方分布。也可以得到相似的 X_{N}^2 的二阶修正。在这里可以使用设计效应矩阵估计值$\hat{\mathbf{D}} = n\,\hat{\mathbf{P}}_{0F}^{-1}\hat{\mathbf{V}}_F$。

基于设计的沃尔德统计量 X_{des}^2与 X_{P}^2 和 X_{N}^2 的二阶拉奥-斯科特修正，都需要方格比例估计值 $\hat{p}_{jk}$的完整的协方差矩阵估计值$\hat{\mathbf{V}}_{des}$。在二手分析的情形下，并不一定给出这一估计值。但可能给出方格设计效应估计值$\hat{d}_{jk}$，可能还有边缘设计效应估计值$\hat{d}_{j+}$与$\hat{d}_{+k}$。使用这些设计效应估计值，可以得到近似的一阶修正。使用以下的估算的方格设计效应，可以计算皮尔逊统计量 X_{P}^2 的最简单的设计效应均值修正，

$$X_{\mathrm{P}}^2(\hat{d}_\bullet) = X_{\mathrm{P}}^2 / \hat{d}_\bullet, \tag{7.44}$$

其中，$\hat{d}_\bullet = \sum_{j=1}^{r}\sum_{k=1}^{c} \hat{d}_{jk}/(rc)$ 是平均方格实际效应。X_{P}^2 的一阶拉奥-斯科特修正为，

$$X_{\mathrm{P}}^2(\hat{\delta}_\bullet) = X_{\mathrm{P}}^2 / \hat{\delta}_\bullet, \tag{7.45}$$

其中，$\hat{\delta}_\bullet$可以从方格与边缘设计效应中得出，

$$\hat{\delta}_\bullet = \frac{1}{(r-1)(c-1)} \sum_{j=1}^{r}\sum_{k=1}^{c} \frac{\hat{p}_{jk}(1-\hat{p}_{jk})}{\hat{p}_{j+}\,\hat{p}_{+k}}\hat{d}_{jk} - \sum_{j=1}^{r}(1-\hat{p}_{j+})\,\hat{d}_{j+} - \sum_{k=1}^{c}(1-\hat{p}_{+k})\,\hat{d}_{+k}$$

而并不需要计算通用设计效应矩阵本身。也可以得到相似的 X_{N}^2 的修正。统计量 $X_{\mathrm{P}}^2(\hat{d}_\bullet)$与 $X_{\mathrm{P}}^2(\hat{\delta}_\bullet)$满足自由度为$(r-1)(c-1)$的卡方分布。$X_{\mathrm{P}}^2(\hat{\delta}_\bullet)$通常比 $X_{\mathrm{P}}^2(\hat{d}_\bullet)$更优。当特征值估计值$\hat{\delta}_l$ 的变动较小时，统计量 $X_{\mathrm{P}}^2(\hat{\delta}_\bullet)$通常较为理想。

当f相对较小时，有不稳定的问题。这时，可以得到 $X_{\mathrm{P}}^2(\hat{\delta}_\bullet)$的 F-校正，

$$FX_{\mathrm{P}}^2(\hat{\delta}_\bullet) = X_{\mathrm{P}}^2(\hat{\delta}_\bullet)/[(r-1)(c-1)], \tag{7.46}$$

它满足自由度为 $(r-1)(c-1)$ 与 f 的 F-分布。一阶修正的内曼统计量 $X_N^2(\hat{\delta}_\bullet)$ 也有相似的校正。

残差分析

当拒绝原独立性假设时,可以计算标准化的基于设计的方格残差,用来详细检查对 H_0 的偏离。残差为,

$$\hat{e}_{jk} = \frac{\hat{F}_{jk}}{\mathrm{s.e}(\hat{F}_{jk})}, \tag{7.47}$$

其中,$\mathrm{s.e}(\hat{F}_{jk})$ 是 $\hat{F}_{jk}$ 的基于设计的标准误估计值,即是,式 7.32 中的相应的方差估计值的平方根。在正的群内相关的情形下,这些基于设计的残差小于在简单随机抽样下相应的残差。可以用 $\mathrm{s.e}_0(\hat{F}_{jk})$ 来代替 $\mathrm{s.e}(\hat{F}_{jk})$ 得到它们,其中,$\mathrm{s.e}_0(\hat{F}_{jk})$ 是 $\hat{F}_{jk}$ 的多项标准误估计值,即是,式 7.33 中的相应的方差估计值的平方根。

范例 7.3

OHC 调查中工作的健康风险与精神压力的独立性检验。我们研究变量 PHYS(工作的身体健康风险:没有或有些)与 PSYCH(分成规模相等的 3 个部分的整体精神压力)是否有联系。注意,两个定类变量组成了类别交叉。表 7.4给出了交叉列表。

表 7.4 OHC **调查中变量** PHYS(**身体健康风险**)**与** PSYCH(**整体心理压力**)**的方格与边缘比例**(**括号内为设计效应估计值**)

PHYS	PSYCH 1	PSYCH 2	PSYCH 3	合 计	n
无	0.227 6 (2.09)	0.218 8 (2.26)	0.207 8 (2.63)	0.654 3 (7.17)	5 130
有	0.116 1 (2.82)	0.104 7 (2.37)	0.125 0 (2.87)	0.345 7 (7.17)	2 711
合计	0.343 7 (1.77)	0.323 6 (1.23)	0.332 7 (1.61)	1.00	
n	2 695	2 537	2 609		7 841

独立性的假设为,$H_0:p_{jk}=p_{j+}p_{+k}$,$j=1,2$,且 $k=1,2,3$,或等价的,$H_0:p_{11}-p_{1+}p_{+1}=0$ 及 $p_{12}-p_{1+}p_{+2}=0$。方格比例的设计效应估计值显示了较大的整群效应。这是由于变量 PHYS 较强的群内相关引起的。这也可以从相应的边缘

设计效应估计值 deff = 7.17 看出来。这样较大的设计效应有一个很自然的解释：相互区隔的单位在物理工作环境上倾向于内部同质，但不同工业行业的工作地点工作环境的差异则较大。另一方面，变量 PSYCH 的边缘设计效应仅仅为中等程度。这也容易理解：心理症状并不能被认为是一个与具体工作强烈相关的现象。方格设计效应估计值的均值为 2.51，也比较大。因此，有效的检验统计量调整整群效应显得很重要。

对于式 7.35、式 7.38 与式 7.39 中的检验统计量，需要相应的残差差异的协方差矩阵估计值$\hat{\mathbf{V}}_F$，$\hat{\mathbf{P}}_{0F}$与$\hat{\mathbf{P}}_F$。

从技术上讲，在计算这些估计值时，使用了完整的偏微分$(rc) \times (rc)$估计值$\hat{\mathbf{H}}$与相应的完整的协方差矩阵估计值$\hat{\mathbf{V}}_{des}$，$\hat{\mathbf{P}}_0$ 与$\hat{\mathbf{P}}$。但在构建检验统计量时，仅使用了这些矩阵的$(r-1)(c-1) \times (r-1)(c-1)$子矩阵。对于 2×3 的表格，我们计算 6×6 的完整表格，但使用 2×2 的子矩阵。使用线性化方法，首先得到一个 6×6 的协方差矩阵估计值$\hat{\mathbf{V}}_{des}$。它是，

$$\hat{\mathbf{V}}_{des} = 10^{-5}\begin{bmatrix} 4.6922 & 0.3207 & 0.6599 & -1.6442 & -1.6965 & -2.3321 \\ 0.3207 & 4.9264 & 1.7922 & -2.5751 & -2.1611 & -2.3030 \\ 0.6599 & 1.7922 & 5.5279 & -2.8972 & -2.5938 & -2.4890 \\ -1.6442 & -2.5751 & -2.8972 & 3.6938 & 1.9619 & 1.4608 \\ -1.6965 & -2.1611 & -2.5938 & 1.9619 & 2.8332 & 1.6562 \\ -2.3321 & -2.3030 & -2.4890 & 1.4608 & 1.6562 & 4.0072 \end{bmatrix}。$$

除了$\hat{\mathbf{V}}_{des}$，为了得到残差差异向量$\hat{\mathbf{F}}$的协方差矩阵估计值$\hat{\mathbf{V}}_F = \hat{\mathbf{H}}'\hat{\mathbf{V}}_{des}\hat{\mathbf{H}}$，需要计算偏微分矩阵$\hat{\mathbf{H}}$。在构建沃尔德统计量时，我们使用 2×1 的残差差异向量，

$$\hat{\mathbf{F}} = \begin{bmatrix} \hat{F}_{11} \\ \hat{F}_{12} \end{bmatrix} = \begin{bmatrix} \hat{p}_{11} - \hat{p}_{1+}\hat{p}_{+1} \\ \hat{p}_{12} - \hat{p}_{1+}\hat{p}_{+2} \end{bmatrix} = 10^{-3}\begin{bmatrix} 2.778 \\ 7.162 \end{bmatrix}。$$

而计算出的相应的完整的$\hat{\mathbf{V}}_F$ 的 2×2 子矩阵为，

$$\hat{\mathbf{V}}_F = 10^{-6}\begin{bmatrix} 7.8147 & -2.8281 \\ -2.8281 & 6.3930 \end{bmatrix}。$$

对于基于设计的沃尔德统计量 $X^2_{des} = \hat{\mathbf{F}}'\hat{\mathbf{V}}_F^{-1}\hat{\mathbf{F}}$，我们得到观测值 $X^2_{des} = 13.41$。它满足自由度为 2 的卡方分布，p-值为 0.0012，在 0.01 水平下显著。式 7.36 与式 7.37 的 X^2_{des}的 F-校正得出观测值 $F_{1.des} = 6.68$，符合自由度为 2 与 244 的 F-分布，p-值为 0.0015；以及 $F_{2.des} = 6.71$，符合自由度为 2 与 245 的 F-分布，p-值与上面相同。F-校正值并没有显著提高未校正的 X^2_{des}。

对于另外的皮尔逊与内曼统计量 X^2_P与 X^2_N 的基于二阶修正的渐进有效检验，我们首先计算式 7.43 中的估算的通用设计效应矩阵如下：

$$\hat{\mathbf{D}} = n\,\hat{\mathbf{P}}_{0F}^{-1}\,\hat{\mathbf{V}}_F = \begin{bmatrix} 1.307\,61 & 0.216\,51 \\ 0.086\,16 & 1.056\,28 \end{bmatrix}。$$

一阶修正因子为$\hat{\delta}_{\bullet} = \mathrm{tr}(\hat{\mathbf{D}})/2 = 1.182$，而特征值的平方和为$\sum_{l=1}^{2}\hat{\delta}_l = \mathrm{tr}(\hat{\mathbf{D}}^2) = 2.863$，得出二阶修正因子$(1+\hat{a}^2) = 1.025$。这些数字表明，平均而言特征值接近于 1，其变化也可忽略不计。

对于等式 7.38 与等式 7.39 中的未修正检验统计量，得到观测值 $X_{\mathrm{P}}^2 = 16.40$与 $X_{\mathrm{N}}^2 = 16.59$。两者均满足自由度为 2 的卡方分布，得到p-值 0.000 3，均在 0.001 水平下显著。注意，与 X_{des}^2相比，X_{P}^2 与 X_{N}^2 要松散一些。对于等式 7.42中的二阶拉奥-斯科特修正，我们得到观测值 $X_{\mathrm{P}}^2(\hat{\delta}_{\bullet},\hat{a}^2) = 13.68$，满足萨特斯韦特调整自由度为 $\mathrm{df}_S = 1.952$ 的卡方分布，p-值为 0.001 0。相较于基于设计的沃尔德统计量，这一检验显得更为松散。但看起来，它在 OHC 调查的检验情形下也较为合理(参见范例 7.2)。

给定有限的信息，我们使用表 7.4 中的设计效应估计值来计算式 7.44、式 7.45 与式 7.46 中的皮尔逊统计量的一阶修正。与一阶拉奥-斯科特修正 $X_{\mathrm{P}}^2(\hat{\delta}_{\bullet}) = 14.02$，$p$-值为 0.000 9 及其 F-校正 $FX_{\mathrm{P}}^2(\hat{\delta}_{\bullet}) = 7.01$，$p$-值为 0.001 1 相比，设计效应均值修正的 $X_{\mathrm{P}}^2(\hat{d}_{\bullet}) = 6.60$，$p$-值为 0.036 9 显得过于保守。设计效应均值修正的保守，是由于$\hat{d}_{\bullet} = 2.51$ 夸大了真实的特征值均值δ所引起的。使用方格与边缘设计效应估计值计算得来的$\hat{\delta}_{\bullet} = 1.182$，给出了一个更优的估计。当定类变量中有一个群内相关较强时，则提醒不要使用设计效应均值修正。与基于设计的沃尔德统计量与二阶拉奥-斯科特修正相比，F-校正一阶拉奥-斯科特修正非常令人满意。

这些检验意味着拒绝 PHYS 与 PSYCH 独立的原假设。我们最后使用式 7.47计算基于设计的标准化方格残差：

PSYCH	PHYS	
	无$\hat{e}_{1k}$	有$\hat{e}_{2k}$
1	0.99	-0.99
2	2.83	-2.83
3	-3.40	3.40

残差分析显示，最大的偏离发生在 PSYCH 的最后一个类别，其方向倾向于那些工作中面临身体健康风险的人。这些类别的标准化残差超过了 $N(0,1)$分布 0.1%临界值 2.58。注意，两个类别的残差和为 0。

与范例 7.2 相同，在检验中，由于样本整群数目较大(250)，基于设计的

沃尔德统计量表现合理。因此,我们认为在检验 PHYS 与 PSYCH 的独立性假设中,沃尔德检验给出了一个满意的方法。当仅给出方格与边缘设计效应时,我们选择皮尔逊(或是内曼)统计量的一阶拉奥-斯科特修正的 F-校正。但当仅给出方格设计效应而没有边缘设计效应时,由于在这一例子中的设计效应均值修正明显过于保守,要得到近似有效的检验方法比较困难。

二维表格中的独立性检验,也可以通过检验两个定类变量的对数线性模型中交互作用的缺损来获得。通过拟合完整对数线性模型 INTERCEPT + PHYS + PSYCH + PHYS * PSYCH,并检查 PHYS 与 PSYCH 交互作用——即 PHYS * PSYCH 项——的沃尔德检验的显著性,来得到独立性检验。基于设计的沃尔德统计量的观测值为 $X_{des}^2 = 13.83$,p-值为 0.001 2,与前面的结果相符。

7.6 本章小结与更多的文献

小　结

在复杂调查中表格的拟合度检验与同质性和独立性检验中,有检验方法来合理修正抽样设计的复杂性。这些复杂性包括个案加权,以获得一致的估算比例,以及整群设计造成的、通常为正的群内相关。普遍意义上,有效的检验方法包括基于设计的沃尔德检验,以及皮尔逊与内曼检验统计量的二阶修正。

在大样本及样本整群数目巨大时,基于设计的沃尔德检验令人满意。在 OHC 调查中,就是如此。沃尔德检验的缺点之一是,它对样本整群数目较小的小样本较为敏感,并导致意料之外的过大的检验结果。MFH 调查是这种设计的一个例子。得出 F 类型校正值的沃尔德统计量的自由度纠正,可以用来解决可能的不稳定的问题。皮尔逊与内曼检验统计量的二阶拉奥-斯科特修正对于不稳定问题并不特别敏感。这一修正在 OHC 与 MFH 调查中表现良好。

基于设计的沃尔德检验与皮尔逊和内曼检验统计量的二阶拉奥-斯科特修正,需要完整的基于设计的协方差矩阵估计值。在对已公开表格的二手分析中,没有给出这样的协方差矩阵估计值,因而仅有近似有效的一阶检验方法。当仅给出方格设计效应估计值时,可以使用设计效应均值修正。但是,如 OHC 调查的例子,这种修正可能过于保守。一阶拉奥-斯科特修正要优于设计效应均值修正。并且,正如 MFH 调查的例子,使用 F-校正,一阶修正在某些情形下可以解决可能的不稳定问题。

由于同质性检验也可以当成一个对数模型的简单应用,以及独立性检验可以当成一个对数线性模型的应用,也可以使用这些在分析复杂调查的计算机软件含有的建模方法。读者可以参考本书的扩展网页,得到更多的这些方法的训练。

在假设检验中,讨论了有限总体方格比例的向量。但是,当有限总体较大时,这些比例接近于无限超总体的相应的方格概率。有限总体可以被看成是这一超总体的单一例子。所以,这里讨论的基于设计的推论,也可以是对无限总体中参数的推论的组成部分。

更多的文献

调查分析文献关注一维与二维频次表分析。霍尔特等(Holt et al. ,1980)与拉奥和斯科特(Rao and Scott,1981,1984,1987)的文章涵盖了 1980 年代的重要理论发展。希迪罗格洛与拉奥(Hidiroglou and Rao,1987a,1987b)及拉奥与托马斯(Rao and Thomas,1988,1989)则给出更多的应用材料。托马斯等(Thomas et al. ,1996)讨论了在复杂抽样下,二维表格的各种独立性检验方法。

这一论题也有其他概论性与专门性材料,如弗里曼与内森在《统计学手册》(Freeman and Nathan,第 6 卷,1988)中的文章,以及桑德尔等(Särndal et al. ,1992)与洛尔(Lohr,1999)的小节。拉奥与托马斯(Rao and Thomas,1988)、斯金纳等(Skinner et al. ,1989)讨论了基于设计与基于模型的推论的双重性。拉奥与托马斯(Rao and Thomas,2003)总结了近来在复杂调查中关于分析定类变量的许多发现。

第8章 多变量调查分析

Multivariate Survey Analysis

多变量方法是分析复杂调查数据的有力工具。本章讨论的多变量分析的情形是，一个回应变量与一组预测或解释变量。对于这样的情形，广泛使用对数模型与线性模型。在复杂抽样设计的条件下，有合适的方法在拟合模型时考虑到了回应变量的群内相关。调查分析的计算机软件也应用了这些方法。与二维表格分析一样，在复杂调查中构建对数与线性模型时，对于估算与检验中剔除整群效应是相当重要的。本章的重点在于讨论较新的方法，并辅以数字化的例子。章节 8.1 与 8.2 介绍多变量方法的范围，以及基本的对数与线性模型。章节 8.2 也讲解了多变量分析中基于设计的其他分析选择。章节 8.3 讨论与演示基于设计的定类变量的分析。对数回归与线性回归分析在章节 8.4 中得以讲解。章节 8.5 给出小结。作为复杂调查的例子，职业健康保健（OHC）调查数据将被用于实际例子中。本书的扩展网页有对例子的进一步讲解。

8.1 方法的范围

拟合多变量模型的目的是，找出回应变量系统性变动的一个科学意义上有趣并简洁的解释。达到这一目的的方式是，利用调查数据，使用一组合理的预测变量对这些变动建立模型。例如，在对象为家庭户的整群样本的健康调查中，为了在发起的健康推动计划中找出针对性的潜在的高风险人口子群，需要研究健康状况与健康服务的使用情况。使用的预测变量，包括样本家庭户的某些社会经济因素，以及家庭成员的人口与行为特征。在一个以教学群体为对象的整群抽样调查中，可能想要研究老师与学生对于学习差异的效应。还有，同样根据以工业公司为整群抽样单位的、与健康相关的工作环境调查数据，可以研究感受到的精神（心理与神经）压力与某些物理和其他工作环境之间的联系。在所有这些调查中，都使用整群抽样来获得样本，但推论则是在个

人层次。更普遍地讲，是超总体框架下的个人层次变量间的关系。

这些调查中的回应变量是二分的（是否患有慢性病；高度或是没有精神紧张）、多项的（学习结果差、中或是好），或是定量或连续的（医生访问次数；精神紧张的主成分分值）。二分与多项的变量对数模型和连续变量的线性模型为这些例子提供了常用的方法。正如将要在调查例子中所看到的，当使用了整群抽样时，回应变量面临群内相关的问题。下面导入的例子简要地讨论群内相关的后果。

导入的例子

让我们更细致地考虑二分与连续回应变量的例子。对于定类预测变量，二分回应变量的数据可以得出一个比例表格；对于连续回应变量，可以得出一个均值表格。从 OHC 调查中，我们有以下感受到的整体精神紧张（PSYCH）变量——最初是从一组精神症状变量中得到的第一个主成分分值的连续变量。对于一个二分回应变量，变量 PSYCH 重新编码，数值 0 表示低于平均值（低紧张组别），数值 1 表示高于平均值（高度紧张组别）。在表格中，我们有 3 个定类预测变量，各有两个类别：应答者的性别与年龄，以及变量 PHYS（身体健康风险）——测量物理工作环境，数值 1 表示更高的工作风险。预测变量的类别交叉生成了组群，它们横切样本整群。主要的兴趣在于精神紧张与物理工作环境的关系。在范例 7.3 中，虽然情形有些不同，PSYCH 含有 3 个类别，但指出了这两个变量间统计上显著的相关。

整个样本中，感觉到高于平均水平的精神压力的百分比当然是 50%，在风险组（PHYS = 1）的比例为 52.2%，即，略微地高出另一个组的百分比。但是，观察表 8.1 中百分比估计值的变动情况，有些子群感觉精神压力的比例要高些。对于两个性别，这一比例随年龄增加而升高；给定年龄组，这一比例在物理环境风险高的组更高。年龄与物理工作环境间可能也有交互作用。

因此，二分回应变量的比例的变动相当有逻辑性。显然，相应的连续性的精神压力的均值也有相似的变动模式。分析组群比例时，选用对数分析；线性分析则适用于组群均值。因为预测变量为定类的，两个例子中均选用方差分析模型。如果数据通过简单随机抽样（SRS）得到，从技术上看，分析是常规性的：从任一商用软件中找出二分对数模型与线性分析的方差分析（ANOVA），找出拟合度高并且简洁的对数与线性模型，得出结论。

但 OHC 调查是以工作单位作为整群的整群抽样。与范例 7.3 相同，回应变量 PSYCH 中的群内相关为正。这一相关将影响分析。如果忽略它，可能会得出错误的结论。从表 8.1 中，可以看出，基于设计的比例估计值都大于 1。整体上的设计效应估计值为 deff = 1.7，表示较大的整群效应。为了得到恰当的分析，应当考虑到这一整群效应。范例 8.1 将显示，除了忽略整群效应，

对于 PSYCH 比例的变动,可以得到一个更简单的模型。

表 8.1 高心理压力组的比例(%)与心理压力连续变量的均值及它们的设计效应估计值,分性别、年龄及应答者物理工作环境(OHC 调查;n =7 841 雇员)

				PSYCH (二分)		PSYCH (连续)	
组群	SEX	AGE	PHYS	%	deff	均值	deff
1	男	-44	0	41.9	1.16	-0.193	1.14
2			1	47.2	1.33	-0.084	1.36
3		45 -	0	46.1	0.87	-0.075	1.05
4			1	52.0	1.18	0.139	1.25
5	女	-44	0	54.1	1.23	0.065	1.61
6			1	62.0	1.38	0.264	1.46
7		45 -	0	53.2	1.65	0.098	1.74
8			1	70.0	1.47	0.656	1.44
合计				50.0	1.69	0.000	1.97

两种主要方法

对于诸如 PSYCH 的群内相关的回应变量,主要有两种合适的多变量分析方法。当群内相关被当成是干扰时,可以设法在估算与检验中剔除它,就像第 7 章所做的。干扰方法(nuisance approach)包括各种对数与线性建模,并主要在调查抽样中发展了很长时间。这种方法有时被称为聚合方法(aggregated approach)。在第 8 章中,我们将讨论用于复杂调查数据中拟合对数与线性模型的常见方法。这些方法是基于最小二乘法(LS)估算与极大似然(ML)估算的变种。

另一方面,当整群作为总体的结构特征而相当重要时,可以用恰当的模型来研究它。这一方法在通用的结构分层级的数据中的多层模型的框架下得到了发展。多层模型也可以用于整群设计中相关回应变量的多变量分析。但是,在复杂调查中,干扰方法有着压倒性的优势,它也是这里讨论的重点。另一种方法,可以被称为分解方法(disaggregated approach),将在本章作简要讨论,而在第 9 章做演示。

估算方法

在对群内相关回应变量建模时,有其他渐进有效的估算方法。对于二分或是多值定类回应变量,当数据如表 8.1 中的多维表格时,我们使用通用最小二乘法(generalized least squares,GLS)的变种来估算。使用 GLS 于复杂调查中时,估算方程将元素权重纳入其中。我们称之为通用加权最小二乘法(generalized weighted least,GWLS)。在章节 8.3 中,我们将讨论用于定类数据的对

数与线性模型的这种非叠代方法。由格里兹尔等(Grizzle et al. ,1969)与科克等(Koch et al. ,1975)引入的 GWLS 方法,可用于线性、对数与指数的比例混合函数。因此,除了对数与线性模型,对数线性模型也将涵盖于其中。

对于复杂调查中二分、多值及计数回应变量模型拟合,一个广泛使用的方法是基于极大似然估算的变种。这样的方法可以纳入元素权重到估算方程中。这种被称为类似然(pseudolikelihood,PML)估算的方法将在章节 8.4 中的二分回应变量的对数分析中加以讨论。在连续变量的线性模型中,将使用纳入元素权重的最小二乘法,这一方法被称为加权最小二乘法(WLS)。在所有这些方法中,当估算回归系数估计值的协方差矩阵时,将使用第 5 章介绍的使用近似的基于设计的方法。线性与非线性模型是更综合的拟合通用线性模型(genera lized linear models)的特例。这引用了内尔德尔与韦德伯恩(Nelder and Wedderburn,1972)以及麦卡洛克与内尔德尔(Mccullagh and Nelder,1987)所讨论的线性、对数与对数线性模型。

第 3 种方法是基于通用估算方程方法(generalized estimating equations, GEE)(Liang and Zeger,梁与齐格尔,1986)。模型的参数用所谓的多变量准似然(multi variate guasilikelihood)方法估算。我们将在章节 8.4 中简要介绍这一方法。与类似然(PML)方法一样,它缘起于通用线性模型方法。

在检验中,基于设计的沃尔德检验统计量与二阶拉奥-斯科特修正检验统计量能够给出渐进有效的检验方法。但是,这些统计量可能会有不稳定的问题,特别是样本整群数目较小时。不稳定性干扰基于设计的沃尔德统计量,导致相对于名义上过大的检验结果及过于复杂的模型。这一特性与二维表格中的沃尔德统计量的问题相似。为了消除不稳定的影响,需要使用诸如 F-校正方法来修正自由度。

虽然这些估算方法有很多共同点,但它们的适用性与特征在某些方面互不相同。为了进一步讨论,我们接下来定义线性与对数模型的主要类型,并且更加正式地介绍相应的模型。

8.2　模型的类型与分析选择

三种类型的模型

在线性模型中,连续回应变量的期望值与预测变量的线性表达式相关。在对数模型中,二分变量的期望值的函数——被称为比率对数或是逻辑斯蒂函数,与预测变量的线性表达式相关。注意,两种模型预测变量的表达式均为线性。但是,基本的差异是,预测变量与回应变量是在线性模型中线性相关

的；而在对数模型中，则假定为非线性的。

为了便于介绍线性与对数模型的类型，分别考虑含有定类预测变量的多维表格与预测变量均是连续性（或者至少有一个是连续性的）的两种情形是有意义的。在这两种情形中，回应变量可以是二分的、多值的、定性的或是连续的。

在诸如表 8.1 的多维表格中，预测变量是类别定性或是类别定量变量。取决于另外的关于它们类型的假设，我们可以得到线性及对数模型的特例。在 ANOVA 类型的模型中，每一个预测变量均被当成定性。性别、职业、社会阶级，以及工业部门是常用的预测变量。对于类别定性变量，可以假定每一个预测变量根据组别单调排序，并将期望取值赋予各个组别。这样，预测变量可以被当成连续的，转而使用回归类型的模型。年龄、收缩血压、家庭月收入，以及精神症状的第一个主成分被分成少数几个组别，成为这样的例子。注意，如范例 7.3 中，原来定量变量的组别也可以当成定性变量。当模型同时含有类别定性与类别定量变量时，我们可以称模型为协方差分析或 ANCOVA 模型。对于 ANOVA 与 ANCOVA 模型，包括交互变量并检验其显著性是常见的。这也是模型构建的一部分。

有时，并不一定要将定量预测变量变成组别，而将数据安排成多维表格形式。这样，我们至少有一个连续预测变量。根据其他预测变量的类型，我们可以得到相应的模型。如果所有预测变量都是连续性的，我们有回归类型模型；额外的定性预测变量得到 ANOVA 模型。应当注意到，在这一例子中，我们用模型表示个人层次的区别；而前一个例子则用模型表示总体子群间的差异。

在连续回应变量的分析中，传统上使用的 ANOVA 与回归分析及 ANCOVA 模型组成了线性模型的特例。我们对于二分与多值回应变量的对数模型使用相似的术语。因此，我们有相应的对数 ANOVA、对数（逻辑斯蒂）回归以及对数（逻辑斯蒂）ANCOVA 模型。

比例的对数与线性模型

由于二分回应变量组群的对数与线性模型在实际中的更为简单且广泛流行，以下的例子与它们相关。下面的对数与线性模型例子中，数据构成一个多维表格，预测变量是含有 u 个组群的互不相交的类别变量，回应变量是二分变量。为了检验回应变量在各个组群中比例的系统变化，可以引入对数与线性模型。这一情形与表 8.1 基本相似。

在一个对数模型中，我们面对的是比例（proportions）p_{j1} 与 p_{j2} 的比率（ratios）的对数形式。其中，前者是成功的比例。我们用 p_j 来表示它，后者因而为 $1-p_j$。通过将未知比例 p_j 的函数 $\log(p_j/(1-p_j))$ 与线性函数 $b_1x_{j1}+b_2x_{j2}+\cdots+b_sx_{js}$ 相连，用模型来表示其变化。这里的“log”表示自然对数。函数

$\log(p_j/(1-p_j))$被称为成功的比率对数。在线性函数中,b_k 是要估算的模型系数,其中的第一个 b_1 是截距项。取值 x_{jk}是对应预测或解释变量 x_k 的,x_1 则给定为 1。其余变量则取决于模型类型。在对数 ANOVA 中,x_k 是表示预测变量的组别的标识变量。在对数回归中,它们是组别的赋值或是原来的连续变量。在对数 ANCOVA 中,x-变量是表示变量与连续变量的混合。系数 b_k 的解释取决于模型的类型以及具体模型中所使用的参数。对数模型的一个优势是,比率比类型(odds-ratio-type)的统计量是现成的。在特例中,也可以解释独立与条件独立的概念。

另一方面,在比例的线性模型中,我们直接面对比例差异(differences of proportions)。因此,总体比例 p_j 与线性函数 $b_1x_{j1} + b_2x_{j2} + \cdots + b_sx_{js}$相连。这一模型构造与对数模型构造相同。模型的某些解释相当容易。但是,独立性以及相连的自然概念则不被解释。

对数与线性模型可以简洁地写成矩阵形式。令 $\mathbf{p} = (p_1, \cdots, p_u)'$为未知组群比例向量,$\mathbf{b} = (b_1, \cdots, b_s)'$为模型系数向量,以及 $\mathbf{X}$ 为 x_{jk}的 $u \times s$ 矩阵,其中的列表示变量 x_k 的取值。通常,$\mathbf{X}$ 被称为模型矩阵(model matrix)。一个假设中的模型可以写成以下形式:

$$F(\mathbf{p}) = \mathbf{Xb}, \tag{8.1}$$

其中,在对数模型的情形下,未知比例向量 $\mathbf{p}$ 的函数向量 $F(\mathbf{p})$为:

$$F(\mathbf{p}) = F(\mathbf{f}(\mathbf{b})) = \log\frac{\mathbf{f}(\mathbf{b})}{1-\mathbf{f}(\mathbf{b})}, \tag{8.2}$$

在线性模型中,由于 F 是单位函数,所以函数向量 $F(\mathbf{p})$等于 $\mathbf{p}$。另外,函数向量 $\mathbf{f}(\mathbf{b})$由以下比率对数函数的倒数推导出来:

$$\mathbf{f}(\mathbf{b}) = F^{-1}(\mathbf{Xb}) = \frac{\exp(\mathbf{Xb})}{1+\exp(\mathbf{Xb})}, \tag{8.3}$$

其中,“exp”表示指数函数。对于线性模型,这一函数向量显然是 $\mathbf{f}(\mathbf{b}) = \mathbf{Xb}$。使用比率对数函数的一个原因是,其值在 0 与 1 之间变动,即与比例 p_j 本身的方位一致。所以,拟合对数模型的预测值总是在范围(0,1)之间。线性模型的构造并没有这一特征。

作为矩阵表达式 8.1 到式 8.3 的演示,让我们考虑以下例子。二分回应变量比例 p_j 的对数与线性 ANOVA 模型有两个二分预测变量 A 与 B。因此,有 4 个组群($u = 4$)及未知比例 p_j 的表格如下:

组群	A	B	p_j
1	1	1	p_1
2	1	2	p_2
3	2	1	p_3
4	2	2	p_4

表格中的变动有 3 个来源：A 的效应，B 的效应，以及 A 与 B 的交互效应。为了涵盖所有变动的来源，模型 $F(\mathbf{p}) = \mathbf{Xb}$ 中共有 4 个系数 b_k。系数 b_1 是截距，b_2 对应 A，b_3 对应 B，而 b_4 对应 A 与 B 的交互效应。这一模型被称为完全模型，通过选择具体矩阵 $\mathbf{X}$，它可以表示为：

$$\begin{bmatrix} F(p_1) \\ F(p_2) \\ F(p_3) \\ F(p_4) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}, \tag{8.4}$$

对于对数模型，函数 $F(p_j)$ 是比率对数，

$$F(p_j) = \text{logit}(p_j) = \log\left(\frac{p_j}{1 - p_j}\right), j = 1,2,3,4,$$

对于线性模型，$F(p_j) = p_j$。在式 8.4 的模型矩阵中，我们首先有一列为 1，用来表示标识变量 x_1。然后，有 3 个比照列，取值为 1 或 -1。其中的第一列对应预测变量 A，即标识变量 x_2；第二列对应预测变量 B，即标识变量 x_3；最后一列对应 A 与 B 的交互效应，即标识变量 x_4。注意，每一个标识变量的参数合计为 0，且每一个预测变量及交互效应有一个标识变量。这是因为预测变量是两组别变量。通常，对于 t 组别的变量，在模型矩阵中，有 $t-1$ 列；对于 t 组别与 v 组别变量的交互效应，有 $(t-1) \times (v-1)$ 列。这与模型的自由度相对应。这些自由度的总和即是模型的系数数目。

刚才使用的参数模式被称为边缘（marginal）或是满秩中心点参数模式。在这种模式下，对于多于两个组别的类别预测变量，每一个标识变量与所有组别的平均值相对比。比如，在一个对数 ANOVA 模型中，系数 b_k 表示在一个比率对数刻度上的不同效应，即是与所有拟合对数模型的均值相对比。在线性 ANOVA 模型中，它们表示未转换刻度上的不同效应，即是与所有拟合比例均值相对比。

在推论中，意识到使用的具体参数模式是很重要的。因为，还有其他常用的参数模式。比如，可以使用一种偏（partial）参数模式或是参照格参数模式。这里，具体的参照组别被假定了，每一个标识变量用来与某一给定的参照组相比较。在这种参数模式下，我们在前面模型矩阵 $\mathbf{X}$ 中的 -1 处用 0 来替代。当可以指出一个确定的参照组时，这一参数形式特别有用。在对数模型中，现在的系数表示与参照组拟合对数模型相比的不同效应。在线性模型中，则表示与参照组拟合比例相比的不同效应。在偏参数化形式中，对数模型中的比率比 $\text{OR}(b_k) = \exp(b_k)$ 的解释就容易了。

在这些参数形式中，我们有函数 $F(p_j)$：

边缘	偏
$F(p_1)=b_1+b_2+b_3+b_4$	$F(p_1)=b_1+b_2+b_3+b_4$
$F(p_2)=b_1+b_2-b_3-b_4$	$F(p_2)=b_1+b_2$
$F(p_3)=b_1-b_2+b_3-b_4$	$F(p_3)=b_1+b_3$
$F(p_4)=b_1-b_2-b_3+b_4$	$F(p_4)=b_1+b_4$

注意,由于两个参数模式中的函数 $F(p_j)$ 必须相等,两个参数模式中相应的系数 b_k 并不相同。比如,边缘参数形式中的系数 b_1 与偏参数化形式中的 b_1 就不等。

到目前为止,我们的讨论限于组群比例的对数与线性 ANOVA 模型。连续回应变量的组群均值的线性 ANOVA 模型的讨论也与此相似。

对于二分回应变量的对数模型、线性回归与 ANCOVA 模型,以及连续回应变量的线性模型,模型矩阵却不同,因而涉及对模型参数的不同解释。

实践中的建模

在拟合一个设定的对数或是线性模型中,主要的任务就是估算模型系数 b_k 以及估算系数的方差。使用估计值,通过检查模型的拟合度来评估模型是否满意,并检验线性假设。在实际中,建模经常涉及使用不同的模型几次重复这一过程。

让我们进一步考虑比例的对数与线性模型。在模型拟合过程中,使用标准的符号,前面的二分变量的 ANOVA 类型的对数模型可以写成为 $F(P)=\log[P/(1-P)]=A+B+A\times B$;线性模型为 $F(P)=P=A+B+A\times B$。3 个模型项对应于预测变量:2 个主效应,一个交互项。这一模型是完全的,因为它包括了所有可能的项;所有模型中,默认包括截距项。在许多线性与对数分析软件中,这样的符号常用于指明模型结构——模型中线性部分的各项。

包括所有可能主效应与交互效应项的完全模型并没有多大意义,因为这一模型有多少自由度就有多少参数。同时,完全模型与数据圆满拟合。但在建模过程中,我们的目标是找出一个简洁的拟合程度较高的模型,使得模型含有尽可能少的项。

使用上面的符号,在这些对数与线性 ANOVA 例子中可能的模型为:

$F(P)=A+B+A\times B$	(完全模型)
$F(P)=A+B$	(主效应模型)
$F(P)=A$	(仅有预测变量 A 的模型)
$F(P)=B$	(仅有预测变量 B 的模型)
$F(P)=\text{INTERCEPT}$	(空心模型)

逐次从模型中排除统计上不显著项,可以得到简洁模型。这一程序对应

于剔除模型矩阵中的列(或是一组列)。通常,拟合度较高的、可供进一步解释的模型介于完全模型与空心模型之间。

组群均值的线性 ANOVA 建模过程与组群比例的对数与线性 ANOVA 建模过程相似。在涉及连续预测变量的对数与线性回归及 ANCOVA 类型模型中,从空心模型开始,连续放入统计上显著或是科学上有意义的项,直到得到合适的模型。应当注意到,不能使用连续预测变量间的交互效应。

在复杂调查中,可以用 GWLS,PML 或是 GEE 方法来估算对数 ANOVA,ANCOVA 或是回归模型的系数。对于回应变量为二分或多值的,而预测变量严格为连续变量的对数回归与 ANCOVA 模型,则使用 PML 或是 GEE 方法。在实际中,在调查分析软件中可以简便地拟合所有这些模型。

在进入使用 GWLS,PML 与 GEE 方法之前,我们详细地讨论复杂调查中多变量分析的特别之处。我们将介绍在不同的抽样设计假设下,正确分析若干选择情形。

分析选择

这里,我们介绍涉及整群、分层、多级抽样与不可忽视的无应答的复杂调查的多变量分析的一些分析选择。面临这样的复杂性,对于选中概率不等以及无应答,可以通过恰当地加权,并恰当地考虑到群内相关而得到模型系数的一致估计值与方差,以及有效的检验结果。

有 3 种具体的分析选择:基于设计的分析选择,以及假定放回式或是无放回式简单随机抽样(SRS)的分析选择。通常,设定放回式。我们称第一种为基于设计的分析选择,它使用了实际上可能很复杂的抽样设计。在基于 SRS 的分析选择中,不管实际使用的抽样设计是否可能更复杂,设定简单随机抽样。第一种基于 SRS 的分析选择纳入了校正不可忽略的无应答的加权。我们称之为加权 SRS 分析选择。第二种 SRS 分析选择被称为未加权 SRS 分析选择。它忽略了包括加权在内的抽样复杂性。

基于设计的分析选择下的分析考虑到了所有抽样复杂性,即是加权、分层与整群。加权 SRS 分析选择忽略分层与整群,而未加权 SRS 分析选择则忽略所有抽样复杂性。在量化设计复杂性对于分析结果的效应时,SRS 分析选择可以用作基于设计的分析选择的参照。

在基于设计的分析选择下,群内相关、不等选中概率以及无应答修正可以完全被考虑到其中。显然,这一分析选择是复杂调查中最适合多变量分析的。所以,调查分析中广泛使用基于设计的分析。本章也将它作为主要分析选择。

取决于抽样设计的特征以及分析用的计算机软件,基于设计的分析选择在实际中的应用有各种途径。抽样设计涉及由于分层与后续分层导致的加权,多级抽样通常需要近似方法简便地拟合基于设计的分析选择。对于从大

量整群的总体中以两级分层整群抽样中得来的数据，这一分析选择的一个简单解决之道是将设计简化为一级分层抽样，并设定以放回式抽取其中的主要抽样单位。在复杂分析性调查中，常见这样的近似方法。使用它需要元素层次的数据。这样的数据包括层级与整群的变量以及权重变量。在第 7 章基于设计的频次表分析中，使用了这样的近似方法。在本章的多变量分析也将使用这一方法。

在基于设计的分析选择的更高级使用中，如有必要，可以在估算时纳入抽样设计的其他特征。这样的例子有，当差异由多级抽样所导致或整群的抽取是无放回式非等概抽取时。这时，假定有以下信息：各级抽样中总体计数、各个主要抽样单位与第一级各层中的每对主要抽样单位的单独或是联合抽中概率的计算。因此，分析程序需要更多的信息。

除了上述的要求以外，如小型芬兰健康（MFH）调查一样，每一层当仅抽取一个主要抽样单位时，基于设计的分析选择的分析可以使用合并层级的手段，重新将样本整群组合成层级。在某些情形下，后续分层的额外加权也是可以的。复杂调查的计算机软件含有许多这样的方法。

在二分回应变量的组群比例多变量分析中，对于基于设计的分析选择，假定了可以计算出一个合适的基于设计的比例的协方差矩阵估计值。在第 5 章中，我们介绍了，基于线性化方法获取一致的协方差估计值。也可以使用诸如折刀法的样本再使用方法。在面临交叉类别或是混合类别的情形时，由于不同组群的比例间的相关可能并不等于 0，这一估计值可以不是对角形的。但是，面临的是相互区隔的类别时，可以假定不同组群的比例间的相关等于 0。这是因为，给定整群的所有元素均落入同一组群。在这种情形下，基于设计的协方差估计值简化为一个对角矩阵。

基于 SRS 的分析选择假定了一个二项的组群比例的协方差矩阵。根据定义，它是对角形的。这一分析选择的有效性取决于实际的抽样设计与组群结构。

基于 SRS 的分析选择假定了放回式简单随机抽样。在加权 SRS 分析选择下，假定使用合适的元素权重可以一致地估算组群比例，并假定这些比例的协方差矩阵为二项的。在未加权分析选择下，也假定了放回式简单随机抽样，并假定数据为自我加权。因此，所有抽样设计的复杂性均被忽略了。

因为这两种基于 SRS 的分析选择，在涉及整群的复杂调查中并不有效，我们将其当成基于设计的分析选择的参照假设，并用于构建合适的通用的设计效应矩阵。在评估整群效应对于多变量分析结果的影响大小时，使用加权的 SRS 分析选择；在检验所有抽样设计的复杂性对于分析结果的影响——包括加权过程的影响时，则使用未加权的 SRS 分析选择作为基于设计的分析选择的参照。

根据抽样设计的分析选择的小结如下：

假　设	允许加权	允许分层	允许整群
基于设计	是	是	是
加权 SRS	是	否	否
未加权 SRS	否	否	否

应当注意到，与二维表格分析一样，在多变量分析中，当有限总体很大时，基于设计的推论方法组成相应的超总体模型参数的推论（Rao and Thomas，1988）。

8.3　定类数据的分析

通用加权最小二乘法估算（GWLS）为组群比例的 ANOVA 类型的对数与线性模型的定类数据分析给出了一种简单的技术。允许包括分层、整群与加权在内的所有抽样设计的复杂性，基于设计的假设则给出了有效范围更大的 GWLS 分析。假定简单随机抽样的加权与未加权 SRS 分析选择分析，可以成为考察整群与加权对于结果的影响作用时的参照。

GWLS 方法在计算上较为简便，因为它对于比例的对数与线性模型并不是叠代计算的。另外的对数模型的类似然（PML）与通用估算方程（GEE）则使用叠代方法，需要更强的计算能力。对于连续预测变量的对数回归，可以使用 PML 与 GEE 方法，而 GWLS 方法则是不合适的。因此，GWLS 方法的适用性比 PML 与 GEE 方法更为有限。

在大样本调查中，任何一种方法得到的结果都相去不远。但在拟合 ANOVA 类型的模型时，由于模型中有多个类别的预测变量，组群的数目也较大，因此需要较大的元素层次的样本来保证每一个组群中含有合理数目的观察个案。这一点对于 GWLS 方法尤为重要。它通常用于样本数目成千上万的类似 OHC 与 MFH 的大规模调查中。为了合理使用 GWLS，PML 与 GEE 方法，大数量的样本整群是有帮助的。回忆一下，在 OHC 调查中就有这一特征。

我们讨论二分回应变量与一组类别预测变量的 GWLS 方法。数据被安排成类似表 8.1 的多维表格。其中，交叉分组构成了 u 个组群，并在各个组群内估算二分回应变量的比例 p_j。在基于设计与加权 SRS 分析选择下的一致估计值 $\hat{p}_j$ 是加权比率类型估计值，其形式为 $\hat{p}_j = \hat{n}_{j1}/\hat{n}_j$，$\hat{n}_{j1}$是组群 j 中二分回应变量的加权样本和，$\hat{n}_j$ 是加权的组群样本规模。在未加权 SRS 分析选择下的未加权比例估计值 $\hat{p}_j^U$，则是使用未加权的 n_{j1} 与 n_j 得到的。

在某一分析选择下，将 GWLS 方法应用于对数与线性模型时，起点是计算相应的比例估计值向量及其协方差矩阵估计值。使用这些估计值，可以估算

模型系数及其协方差矩阵；进一步，可以估算拟合比例及其协方差矩阵。另外，还可以进行模型拟合度的沃尔德检验，以及模型系数的线性假设的沃尔德检验。最后，更进一步进行所选中模型的拟合度残差分析。

基于设计的 GWLS 估算

在基于设计的分析选择下，模型 $F(\mathbf{p}) = \mathbf{Xb}$ 的 $s \times 1$ 的模型系数向量 $\mathbf{b}$ 的一致 GWLS 估计值$\hat{\mathbf{b}}_{des}$——本小节中用$\hat{\mathbf{b}}$表示——为，

$$\hat{\mathbf{b}} = (\mathbf{X}'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}F(\hat{\mathbf{p}}), \tag{8.5}$$

其中，$\hat{\mathbf{V}}_{des}$是一致组群比例估计向量 $\hat{\mathbf{p}}$ 的协方差矩阵的一致估计值，$\mathbf{H}\,\hat{\mathbf{V}}_{des}\mathbf{H}$ 是函数向量 $F(\hat{\mathbf{p}})$的协方差矩阵估计值。估计值 $\hat{\mathbf{V}}_{des}$可以——举例而言——通过第 5 章描述的线性化方法得到。因此，GWLS 估算方程（式 8.5）是基于一致估算函数 $F(\hat{p}_j)$及其基于设计的协方差矩阵估计值。这些方程也显示，在得到$\hat{b}_k$ 的过程中，不需要叠代计算。“GWLS”名称的来源之一是，在得到比例向量估计值及其协方差矩阵估计值中，GLS 估算方程使用了元素权重。

式 8.5 中的 GWLS 估计值 $\hat{\mathbf{b}}$，适用于组群比例的对数与线性模型。但，函数向量的协方差矩阵估计值中的 $\mathbf{H}$ 矩阵在两种模型中却不同。在对数模型中，函数 $F(\hat{p}_j)$的偏微分 $u \times u$ 对角矩阵 $\mathbf{H}$，其对角元素的形式为 $h_j = 1/[\hat{p}_j(1-\hat{p}_j)]$。在线性模型中，矩阵 $\mathbf{H}$ 是一个单位矩阵，主对角线为 1，其余为 0。

在一个对数 ANOVA 模型的偏参数化下（参见章节 8.2），其中的对应与预测变量组别的模型矩阵 $\mathbf{X}$ 的列是二分变量，而估计值$\hat{b}_k$ 的解释可以使用对数比率比。因此，估计值 $\exp(\hat{b}_k)$是在控制模型中其他项的效应下相应组别相对于参照组别的比率比。模型系数估计值的这种解释常见于传染病学与社会科学。

在得到系数的沃尔德检验统计量的过程中，使用了式 8.5 中模型系数估计值$\hat{b}_k$ 的协方差矩阵估计值 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$。这一 $s \times s$ 的协方差矩阵为，

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}}) = (\mathbf{X}'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}\mathbf{X})^{-1}。 \tag{8.6}$$

适当选择 $\mathbf{H}$，这一估计值适用于对数与线性模型。$\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$的对角元素给出了基于设计的模型系数估计值$\hat{b}_k$ 的方差估计值 $\hat{v}_{des}(\hat{b}_k)$，它被用来获取相应的标准误估计值 $s.e(\hat{b}_k) = \hat{v}_{des}{}^{1/2}(\hat{b}_k)$。在对数模型中，使用这些标准误估计值，可以计算以下比率比 $\exp(\hat{b}_k)$近似的 95% 置信区间：

$$\exp(\hat{b}_K \pm 1.96 \times s.e_{des}(\hat{b}_k))。 \tag{8.7}$$

在实际中，另外两个协方差矩阵估计值非常有用。它们是拟合比率对数

向量 $\hat{\mathbf{F}} = \mathbf{X}\hat{\mathbf{b}}$的 $u \times u$ 的协方差矩阵估计值 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{F}})$，以及拟合比例向量 $\hat{\mathbf{f}} = F^{-1}(\mathbf{X}\hat{\mathbf{b}})$的协方差矩阵估计值 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{f}})$。它们的公式为，

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{F}}) = \mathbf{X}\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})\mathbf{X}' \tag{8.8}$$

与

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{f}}) = \hat{\mathbf{H}}^{-1}\hat{\mathbf{V}}_{des}(\hat{\mathbf{F}})\hat{\mathbf{H}}^{-1}。 \tag{8.9}$$

在线性模型中，这些协方差矩阵显然相同，因为拟合函数与拟合比例相等。在对数模型中，对角矩阵 $\hat{\mathbf{H}}$ 的对角元素为 $\hat{h}_j = 1/[\hat{f}_j(1-\hat{f}_j)]$，$\hat{f}_j = f_j(\hat{\mathbf{b}})$是由以下方程计算出的拟合比例向量 $\hat{\mathbf{f}}$ 的元素，

$$\hat{\mathbf{f}} = \mathbf{f}(\hat{\mathbf{b}}) = \exp(\mathbf{x}\hat{\mathbf{b}})/(1+\exp(\mathbf{X}\hat{\mathbf{b}}))。 \tag{8.10}$$

为了得到拟合函数与拟合比例的基于设计的标准误，需要协方差矩阵估计值（式 8.8 与式 8.9）的对角元素。

拟合度与相关检验

检查模型的拟合度是组群比例的对数与线性建模过程的重要部分。将总和协变（总和卡方，total chi-square）分解成模型的协变（模型卡方，model chi-square）与残差协变（残差卡方，residlhal chi-square），可以得到各种拟合度统计量。因此，我们有，

总和卡方 = 模型卡方 + 残差卡方

它有与线性回归和 ANOVA 相似的总和方差的分解。通常使用度量残差协变的、基于设计的沃尔德检验统计量 X^2_{des}作为模型拟合度的指标。这一统计量为，

$$X^2_{des} = (F(\hat{\mathbf{p}}) - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}(F(\hat{\mathbf{p}}) - \mathbf{X}\hat{\mathbf{b}}), \tag{8.11}$$

在基于设计的分析选择下，它满足自由度为 $u-s$ 的渐进卡方分布。相对于残差自由度，这一统计量的较小值表示模型拟合度高。显然，完全模型的拟合度是完美的。度量模型整体协变的、用 $X^2_{des}(overall)$ 表示的沃尔德统计量，用来检验所有模型系数均为 0 的假设。它的形式为，

$$X^2_{des}(overall) = F(\hat{\mathbf{p}})'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}F(\hat{\mathbf{p}}) - X^2_{des}, \tag{8.12}$$

其中，第一个平方项度量整体协变，而第二个平方项是讨论的模型的式 8.11 中的残差卡方。这一统计量满足自由度为 s 的渐进卡方分布。同样的，用 $X^2_{des}(gof)$ 表示的沃尔德统计量可以用来检验除了截距项的所有模型系数均为 0 的假设。这个统计量被定义为，只含有截距的模型与含有所有项的模型的式 8.11中的残差卡方统计量观测值之间的差。因此，它满足自由度为$s-1$的渐进卡方分布。有时，统计量 $X^2_{des}(overall)$被称为整体模型检验，而$X^2_{des}(gof)$为拟合度检验。注意，所有这些统计量都适用于组群比例的对数与线性模型。

使用以下沃尔德统计量，可以检验模型系数向量 $\mathbf{b}$ 的线性假设(linear hypotheses) $H_0:\mathbf{Cb}=\mathbf{0}$,

$$X^2_{des}(\mathbf{b}) = (\mathbf{C}\hat{\mathbf{b}})'(\mathbf{C}\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})\mathbf{C}')^{-1}(\mathbf{C}\hat{\mathbf{b}}), \tag{8.13}$$

其中，$\mathbf{C}$ 是 $c\times s(c\leqslant s)$ 的参照编码矩阵。在基于设计的分析选择下，这一统计量满足自由度为 c 的渐进卡方分布。这个统计量用于在使用以下沃尔德统计量来检验模型单一参数假设 $H_0:b_k=0$ 时，

$$X^2_{des}(b_k) = \hat{b}_k^2/\hat{v}_{des}(\hat{b}_k), \qquad k=1,\cdots,s,$$

它是自由度为 1 的渐进卡方分布。注意，对于相应的 t-检验，等式 $t^2_{des}(b_k) = X^2_{des}(b_k)$ 成立。

另一个在渐进意义上有效的、模型参数的线性假设的检验程序，是使用萨特斯韦特方法对二项沃尔德检验统计量的二阶拉奥-斯科特修正。这样的技术与第 7 章用于皮尔逊与内曼检验统计量的相似。在式 8.5 中用 $\hat{\mathbf{p}}$ 的二项协方差矩阵估计值 $\hat{\mathbf{V}}_{bin}$ 代替 $\hat{\mathbf{V}}_{des}$，我们首先计算 GWLS 估计值 $\hat{\mathbf{b}} = \hat{\mathbf{b}}_{bin}$，并构建相应的沃尔德检验统计量 $X_{bin}{}^2(\mathbf{b})$ 如下：

$$X^2_{bin}(\mathbf{b}) = (\mathbf{C}\hat{\mathbf{b}})'(\mathbf{C}\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}})\mathbf{C}')^{-1}(\mathbf{C}\hat{\mathbf{b}}),$$

其中，$\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}})$ 是通过在式 8.6 中用 $\hat{\mathbf{V}}_{bin}$ 代替 $\hat{\mathbf{V}}_{des}$ 得到的二项 GWLS 估计值的协方差矩阵估计值。二阶修正的沃尔德统计量为，

$$X^2_{bin}(\mathbf{b};\hat{\delta}_{\bullet},\hat{a}^2) = \frac{X^2_{bin}(\mathbf{b})}{\hat{\delta}_{\bullet}(1+\hat{a}^2)}, \tag{8.14}$$

其中，一阶与二阶修正因子 $\hat{\delta}_{\bullet}$ 与 $(1+\hat{a}^2)$ 从以下 $c\times c$ 的通用设计效应矩阵中计算而来，

$$\hat{\mathbf{D}} = (\mathbf{C}\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}})\mathbf{C}')^{-1}(\mathbf{C}\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})\mathbf{C}') \tag{8.15}$$

并使得

$$\hat{\delta}_{\bullet} = \mathrm{tr}(\hat{\mathbf{D}})/c$$

是通用设计效应矩阵估计值的特征值 $\hat{\delta}_k$ 的均值。而

$$1+\hat{a}^2 = \sum_{k=1}^{c}\hat{\delta}_k^2/(c\hat{\delta}_{\bullet}^2),$$

其中的特征值的平方和由以下公式计算得出，

$$\sum_{k=1}^{c}\hat{\delta}_k^2 = \mathrm{tr}(\hat{\mathbf{D}}^2)。$$

在基于设计的分析选择下，二阶修正统计量 $X^2_{bin}(\mathbf{b};\hat{\delta}_{\bullet},\hat{a}^2)$ 满足萨特斯韦特校正自由度为 $\mathrm{df}_S = c/(1+\hat{a}^2)$ 的渐进卡方分布。与检验模型的各个参数一样，当 $c=1$ 时，由于通用设计效应矩阵简化为一个因子，而修正也简化成

一阶，我们有$(1+\hat{a}^2)=1$。在复杂调查的分析软件中含有这个检验统计量。

不稳定的情形

由于拟合度沃尔德统计量 X^2_{des}，$X^2_{des}(overall)$ 与 $X^2_{des}(gof)$，以及模型参数的线性假设统计量 $X^2_{des}(\mathbf{b})$ 在基于设计的分析选择下满足渐进卡方分布，当样本整群数目 m 相对于组群数目 u 较大时，它们是令人满意的。但是，当协方差估计值 $\hat{\mathbf{V}}_{des}$ 不稳定时，检验统计量相对于名义上的显著水平，可能取值较大。当相对于残差与模型自由度，估计值 $\hat{\mathbf{V}}_{des}$ 的自由度 $f=m-H$ 较小时，这样的情形可能发生。

与用于第 7 章的同质性与独立性假设的相似，F-校正的沃尔德检验统计量可以补救不稳定的影响。对于式8.11中的拟合度统计量，自由度校正为，

$$F_{1.des}=\frac{f-(u-s)+1}{f(u-s)}X^2_{des}, \tag{8.16}$$

满足自由度为$(u-s)$与$(f-(u-s)+1)$的 F-分布，以及，

$$F_{2.des}=X^2_{des}/(u-s), \tag{8.17}$$

满足自由度为$(u-s)$与f的 F-分布。使用相应的自由度 s 或者$(s-1)$来代替$(u-s)$，这些 F-校正可以从沃尔德统计量 $X^2_{des}(overall)$ 与 $X^2_{des}(gof)$ 中推导出来。

也可以推导出相似的、模型参数的线性假设的沃尔德统计量的 F-校正。对于式8.13中的统计量，它们是，

$$F_{1.des}(\mathbf{b})=\frac{f-c+1}{fc}X^2_{des}(\mathbf{b}) \tag{8.18}$$

以及

$$F_{2.des}(\mathbf{b})=X^2_{des}(\mathbf{b})/c, \tag{8.19}$$

分别满足自由度为 c 与$(f-c+1)$，以及自由度为 c 与 f 的 F-分布。

二阶拉奥-斯科特修正可以被认为是抗不稳定性问题的。但是，对于式8.14中的二阶修正统计量，也可以推导出 F-校正。它是，

$$F_{bin}(\mathbf{b};\hat{\delta}_{\bullet},\hat{a}^2)=(1+\hat{a}^2)X^2_{bin}(\mathbf{b};\hat{\delta}_{\bullet},\hat{a}^2)/c=X^2_{bin}(\mathbf{b})/(c\hat{\delta}_{\bullet}), \tag{8.20}$$

它满足自由度为 df_S 与 f 的 F-分布。

当 f 较大时，F-校正对于检验的 p-值的影响较小。但是，当 f 相对较小时，特别是当 f 与残差自由度很接近时，校正就有影响。在严重的不稳定情形下，统计量 $F_{1.des}$，$F_{1.des}(\mathbf{b})$ 或是 $F_{1.des}(\mathbf{b};\hat{\delta}_{\bullet},\hat{a}^2)$ 更优。复杂调查的分析软件将这些校正统计量纳入作为检验选项。

残差分析

通过计算原始与标准化的残差,仔细检查选中模型的拟合度是有必要的。这也可以用来发现可能的偏远的组群比例。原始残差就是拟合比例 $\hat{f}_j$ 与相应的观测比例 $\hat{p}_j$ 之间的差($\hat{p}_j - \hat{f}_j$)。在基于设计的分析选择下,为计算标准化残差,首先要得到以下原始残差的协方差矩阵估计值 $\hat{\mathbf{V}}_{res}$,

$$\hat{\mathbf{V}}_{res} = \mathbf{H}^{-1}(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H} - \hat{\mathbf{V}}_{des}(\hat{\mathbf{F}}))\mathbf{H}^{-1}, \tag{8.21}$$

其中,$\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H}$ 与 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{F}})$ 分别是观测函数向量 $F(\hat{\mathbf{p}})$ 与拟合函数向量 $\hat{\mathbf{F}} = \mathbf{X}\hat{\mathbf{b}}$ 基于设计的协方差矩阵估计值。矩阵 $\mathbf{H}$ 则取决于拟合的模型类型——对数或是线性。使用式 8.21,标准化残差计算如下,

$$\hat{e}_j = (\hat{p}_j - \hat{f}_j)/\sqrt{\hat{v}_j}, j = 1, \cdots, u, \tag{8.22}$$

其中,$\hat{v}_j$ 是残差协方差矩阵 $\hat{\mathbf{V}}_{res}$ 的对角元素。较大的标准化残差意味着,模型解释相应组群比例较弱。由于标准化残差是近似的标准正态变量,可以将它与 $N(0, 1)$ 分布的临界值相比较。

设计效应估算

GWLS 方法的一个主要特征是它对于各种模型设定与各种抽样设计的灵活性。在涉及整群与分层的复杂多级抽样设计的基于设计的分析选择下,基于设计的 GWLS 方法看起来是有效的。当选择合适的比例估算公式,并且其协方差估算公式能够反映抽样设计的复杂性时,GWLS 方法也可用于简单的设计中。

在加权 SRS 分析选择下,为得到模型系数相应的 GWLS 估计值 $\hat{\mathbf{b}}$ 与协方差矩阵估计值 $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}})$,在等式 8.5 与等式 8.6 中,使用了一致比例估计值 $\hat{\mathbf{p}}$ 及其二项协方差矩阵 $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}})$。在未加权的 SRS 分析选择下,也是如此。其中,使用了相应的 $\hat{\mathbf{p}}^U$ 与 $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}}^U)$。GWLS 估算方程显示,在基于 SRS 分析选择下得到的估计值 $\hat{\mathbf{b}}_k$ 与基于设计分析选择下得到的在数值上并不相同。

在无法剔除整群对于估算的模型系数的标准误估计值的影响效应时,基于 SRS 的分析选择的受限较大。这一影响效应表现在模型系数估计值的设计效应估计值中。使用模型系数协方差矩阵估计值 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ 与 $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}}^*)$ 的对角元素,可以计算设计效应估计值。所以,我们有,

$$\hat{d}(\hat{b}_k) = \hat{v}_{des}(\hat{b}_k)/\hat{v}_{bin}(\hat{b}_k^*), \quad k = 1, \cdots, s, \tag{8.23}$$

其中,$\hat{b}_k^*$ 表示在加权或是未加权的 SRS 分析选择下得到的模型系数估计值。

在未加权分析选择下，这些设计效应表示所有抽样复杂性的影响；在加权分析选择下，表示整群的影响。计算在两种 SRS 分析选择下的设计效应估计值非常有意义，因为这样可以检查加权对于设计效应的影响。

选择模型的标准

应当选择比例的哪一个模型类型，对数的或是线性的？在某些学科里，一种类型比另一种类型更标准。但是，通常是明确地偏向其中一方是不可能的。看起来，对数模型是有优势的，比如可以——在某些情形下，用标准独立概念来——解释比率比。另外，作为更宽泛的指数族模型的一员，二项比例的对数模型涉及线性模型不具有的简便统计特征。虽然，这些特征并不必要应用于复杂调查中的对数模型，在这种调查中也要注意使用对数模型。

另一方面，对于熟悉连续变量的线性 ANOVA 的，比例的线性模型给出了特别简便的建模方法。在线性刻度上的可加性，线性模型的系数描述了比例本身之间的差异，而非比例对数间的差异。但在实际中，当比例的范围在 0.2 ~ 0.8 时，对数与线性模型估算的模型系数并没有太大差别。在下面的例子中，我们用典型的卫生学科的分析来比较对数与线性模型。

范例 8.1

GWLS 方法的对数与线性 ANOVA。让我们将 GWLS 方法应用于表 8.1 所示的简单 OHC 调查中组群比例的对数与线性建模。我们的目标是，用模型来解释二分回应变量 PSYCH（测量总体上的精神压力）在由性别、年龄以及变量 PHYS（应答者特殊的工作环境）所分隔开的 $u = 8$ 个组群中的协变。表 8.2给出了分析选择更详细的说明。除了组群比例 $\hat{p}_j$、标准误 $s.e_j$ 与设计效应 d_j 以外，也包括了原始样本规模 $\hat{n}_j$ 以及每个组群所包含的样本整群数目 m_j。注意，组群比例在 0.5 左右波动。

表 8.2 高心理压力组比例 $\hat{p}_j$、标准误 s.e$_j$、设计效应估计值 $\hat{d}_j$ 及组群样本规模 $\hat{n}_j$ 与样本整群数 m_j（OHC 调查）

组群 j	SEX	AGE	PHYS	$\hat{p}_j$	s.e$_j$	$\hat{d}_j$	$\hat{n}_j$	m_j
1	男	-44	0	0.419	0.012 8	1.16	1 734	230
2			1	0.472	0.014 5	1.33	1 578	198
3		45 -	0	0.461	0.017 8	0.88	690	186
4			1	0.520	0.024 7	1.18	483	138
5	女	-44	0	0.541	0.012 5	1.23	1 966	240
6			1	0.620	0.027 0	1.38	447	152
7		45 -	0	0.532	0.023 6	1.65	740	185
8			1	0.700	0.039 1	1.48	203	101
合计				0.500	0.007 3	1.69	7 841	250

在这一分析中,基于设计的分析选择给出了有效的对数与线性模型建构。如比例的设计效应估计值的均值大于 1 所显示,抽样设计涉及整群效应。平均设计效应估计值为 1.28。另外,组群由交叉类别所构成,这由每个组群包含了较大数目的样本整群所显示出来。更明显的是,这一特征可以从图8.1显示的组群比例的基于设计的协方差矩阵估计值 $\hat{\mathbf{V}}_{des}$ 中看出来。可以注意到,在协方差矩阵估计值的对角线之外,有非 0 的协方差项。这一估计值看起来相对稳定,因为协方差估计值大大小于相应的方差估计值。$\hat{\mathbf{V}}_{des}$ 的条件数为 12.1,也显示了稳定性。为了便于比较,相应的二项协方差矩阵估计值 $\hat{\mathbf{V}}_{bin}$ 也包括在此。

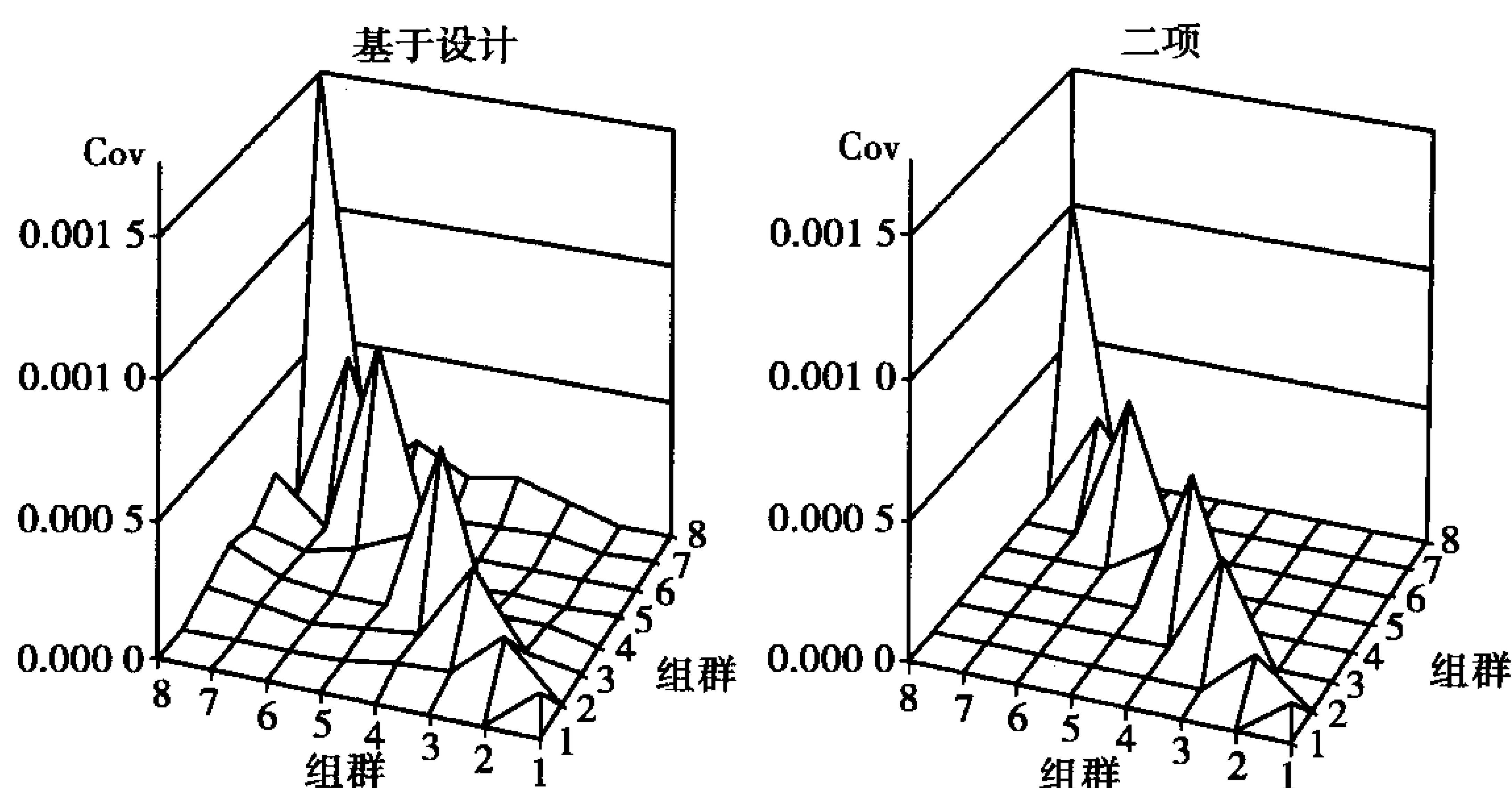


图 8.1 组群比例估计值 $\hat{p}_j$ 的基于设计与二项协方差矩阵估计值

我们讨论在基于设计分析选择下的建模过程,并使用未加权 SRS 分析选择为参照。这里,共有 3 个预测变量。它们的主效应、1 个截距加上 4 个交互效应,共有 8 项出现在完全的对数与线性 ANOVA 模型中。可以写成以下形式:

F(P) = INTERCEPT + SEX + AGE + PHYS + SEX * AGE +
SEX * PHYS + AGE * PHYS + SEX * AGE * PHYS,

其中,对数模型的函数为 $F(P) = \log[P/(1-P)]$,线性模型的函数为 $F(P)=P$,P 表示 PSYCH 取值较大的组别的比例。

在建模过程中,我们首先拟合完全对数与线性模型,并检验 3 个预测变量的交互效应的显著性。如果不显著,我们剔除这一项,并开始检查两个变量间的交互效应,以便进一步简化模型。当得到一个拟合度令人满意的模型时,建模过程便完成了。这种逐步的过程是常见于对数线性与对数 ANOVA 模型中的、所谓的向后排除(backward elimination)。

让我们更仔细地讨论对数模型的拟合结果。在基于设计的分析选择下,主效应模型拟合度令人满意,并且无法进一步简化。表 8.3 给出了模型简化过程。在比较完全模型 5 与模型 4 时,得到了沃尔德统计量 X^2_{des} 的差值。统计

量差值的计算为 $X^2_{des}(overall;\ 5) - X^2_{des}(overall;\ 4) = 78.84 - 76.90 = 1.94$。将它与自由度为 1 的卡方分布比较，得到不显著的 p-值为 0.163 5。因此，交互效应项可以从模型 5 中剔除。主效应模型（模型 1）拟合度的沃尔德统计量的观测值为 $X^2_{des} = 78.84 - 72.39 = 6.45$，其自由度为 4，$p$-值为0.168 1，表示拟合度较为满意。

完全模型得到了相当的简化，建模过程生成了一个仅含有主效应的简单结构模型。所以，SEX 与 PHYS 的交互效应并不显著。在基于 SRS 分析选择下，拟合对数模型时，我们将回到这一结论。

在这里使用的偏参数化中，每一个预测变量的模型系数的第一个组别均被设定为 0。最后组群的第一个组别是参照组群——这里是表 8.2 中的组群 7。在主效应模型中，要估算 4 个系数。事实上，GWLS 估计值$\hat{b}_k$ 由以下模型矩阵得出，

表 8.3 基于设计分析选择下，完整模型的沃尔德统计量（X^2_{des}）与其他简便对数 ANOVA 模型比较的统计量的差

模型	df	整体 X^2_{des}	p-值	模型比较	df	X^2_{des}差	p-值
5	8	78.84	0.000 0	—	1	—	—
4	7	76.90	0.000 0	5 - 4	1	1.94	0.163 5
3	6	76.09	0.000 0	4 - 3	1	0.81	0.369 3
2	5	74.78	0.000 0	3 - 2	1	1.31	0.253 3
1	4	72.39	0.000 0	2 - 1	1	2.39	0.121 8

模型 5：SEX + AGE + PHYS + SEX * AGE + SEX * PHYS + AGE * PHYS + SEX * AGE * PHYS

模型 4：SEX + AGE + PHYS + SEX * AGE + SEX * PHYS + AGE * PHYS

模型 3：SEX + AGE + PHYS + SEX * PHYS + AGE * PHYS

模型 2：SEX + AGE + PHYS + SEX * PHYS

模型 1：SEX + AGE + PHYS

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}。$$

拟合模型可以用$\hat{b}_k$ 及模型矩阵重新写成，

$$F(\hat{f}_j) = \hat{b}_1 + \hat{b}_2(\mathrm{SEX})_j + \hat{b}_3(\mathrm{AGE})_j + \hat{b}_4(\mathrm{PHYS})_j, \quad j = 1, \cdots, 8,$$

其中，对数模型为 $F(\hat{f}_j) = \log[\hat{f}_j/(1-\hat{f}_j)]$，线性模型为 $F(\hat{f}_j) = \hat{f}_j$，标识变量 SEX，AGE 与 PHYS 的取值为模型矩阵 **X** 的第 2，3 与 4 列。

让我们进一步考虑主效应对数模型的估算与检验结果。表 8.4 给出了模型系数的估算结果。

在这个表中，与期望一致，女性与老龄组的估算系数 $\hat{b}_2$ 与 $\hat{b}_3$ 为正，其相应的 t-检验得到了显著的 p-值。控制性别与年龄的、PHYS 中更多工作风险的组别的估算系数 $\hat{b}_4$ 为正，并对应明显显著的 t-检验。应当注意到，这里使用的 t-检验统计量的绝对值等于式 8.19 中 F-校正的沃尔德统计量的平方根。由于整群效应，估算的模型系数的设计效应估计值 $\hat{d}(\hat{b}_k)$ 大于 1。因此，模型系数的二项标准误估计值将小于相应的基于设计的估计值。

表 8.4　整体心理压力的基于设计的对数 ANOVA 估算值（用 GWLS 方法拟合模型）

模型变量	β 系数	设计效应	标准误	t-检验	p-值	比率比	比率比的 95% 置信区间	
							下限	上限
截距	-0.328 2	1.32	0.063 5	-7.02	0.000 0	0.72	0.66	0.79
性别								
男	0	n. a.	0	n. a.	n. a.	1	1	1
女	0.466 3	1.44	0.057 9	8.06	0.000 0	1.59	1.42	1.79
年龄								
-44*	0	n. a.	0	n. a.	n. a.	1	1	1
45 -	0.138 5	1.23	0.057 0	2.43	0.015 9	1.15	1.03	1.28
身体健康风险								
无	0	n. a.	0	n. a.	n. a.	1	1	1
有	0.256 8	1.30	0.057 4	4.48	0.000 0	1.29	1.16	1.45

*：参照组；参数值为 0。

n. a.：不适用。

使用更有意义的、PHYS 中更多工作风险的组别的参数估计值 $\hat{b}_4 = 0.256\,8$，可以根据式 8.7 得到相应的控制性别与年龄的比率比及其 95% 置信区间。比率比（OR）估计值为 $\exp(\hat{b}_4) = 1.29$，其 95% 置信区间计算为，

$$\exp(0.256\,8 \pm 1.96 \times 0.057\,4) = (1.16, 1.45)。$$

控制了性别与年龄后，更多工作风险的组别面临较高精神压力的概率，为更少工作风险组别的 1.3 倍。这一结果与 t-检验结果一致，因为 95% 置信区间没有包括 1——参照组的比率比。

我们现在转向最后主效应 ANOVA 模型中各项的检验结果（表 8.5）。不同的沃尔德检验统计量及其 F-校正构成一组观测值。第 1 组检验统计量对

应于原来的基于设计的沃尔德统计量(式 8.13);第 2 组对应于 F-校正统计量(式 8.18);第 3 组对应于萨特斯韦特修正的二项统计量(式 8.14);最后,第 4 组对应于 F-校正统计量(式 8.20)。基于设计的沃尔德统计量 $X^2_{des}(\mathbf{b})$ 与二阶修正二项统计量 $X^2_{bin}(\mathbf{b};\hat{\delta}_{\bullet},\hat{a}^2)$ 给出相似的结果。因此,基于设计的沃尔德统计量在这里令人满意。这主要是由于协方差矩阵估计值 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ 的稳定性。因为估计值 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ 的自由度较大($f=245$),所以 F-校正对原来检验的 p-值影响不大。

表 8.5 整体心理压力最终对数 ANOVA 模型项检验统计量的观测值与 p-值(用 GWLS 方法拟合模型)

比照	Df	(1)基于设计的沃尔德检验	p-值	(2)对(1)的 F-校正	p-值	(3)对二项沃尔德检验的拉奥-斯科特二阶修正	p-值	(4)对(3)的 F-校正	p-值
SEX	1	64.92	0.000 0	64.92	0.000 0	64.92	0.000 0	64.92	0.000 0
AGE	1	5.90	0.015 1	5.90	0.015 9	5.90	0.015 3	5.90	0.015 9
PHYS	1	20.04	0.000 0	20.04	0.000 0	20.04	0.000 0	20.04	0.000 0

(1)等式 8.13,(2)等式 8.18,(3)等式 8.14,(4)等式 8.20。

虽然在这个分析中,不同的检验统计量的结果之间没有冲突,但在有的情形下,选择合适的统计量相当关键。特别是在样本整群数目 m 较小,且组群数目 u 与 m 相去不远时,更是如此。这时,可以选择某些 F-校正来补偿不稳定的影响。

为了进一步检查模型的拟合度,让我们现在计算残差分析中的拟合比例及原始与标准化残差。表 8.6 给出了这些数据。

表 8.6 基于设计分析选择下,对数 ANOVA 模型中 PSYCH 比例的观测值 $\hat{p}_j$ 与拟合值 $\hat{f}_j$ 及它们的标准误,以及原始与标准化残差 $(\hat{p}_j-\hat{f}_j)$ 与 $\hat{e}_j$

组群	SEX	AGE	PHYS	$\hat{p}_j$	s. e$(\hat{p}_j)$	$\hat{f}_j$	s. e$(\hat{f}_j)$	$(\hat{p}_j-\hat{f}_j)$	$\hat{e}_j$
1	男	-44	0	0.419	0.012 8	0.419	0.011 4	0.000 0	0.000 0
2			1	0.472	0.014 5	0.482	0.012 2	-0.010 0	-1.270
3		45 -	0	0.461	0.017 8	0.453	0.014 2	0.008 2	0.771
4			1	0.520	0.024 7	0.517	0.016 7	0.002 9	0.160
5	女	-44	0	0.541	0.012 5	0.534	0.011 5	0.006 2	1.306
6			1	0.620	0.027 0	0.597	0.016 0	0.022 2	2.012
7		45 -	0	0.532	0.023 6	0.569	0.015 6	-0.036 3	-2.073
8			1	0.700	0.039 1	0.630	0.019 9	0.069 2	1.993

除了最后 3 个组群中较大的原始残差以外,比例的观测值与拟合值很相

近。最后两组的标准化残差超过了 $N(0,1)$ 分布 5% 的临界值 1.96，所以，对这两个组群的模型拟合有问题。应当注意到，拟合比例与残差独立于模型的参数选择。

简要讨论在其他分析选择下的对数分析并将它作为基于设计的分析选择分析结果的参照是有用的。在这里，我们对于描述 SEX 与 PHYS 交互效应的 SEX * PHYS 项的重要性很感兴趣。它在基于设计的分析选择下，并不显著。表 8.7 给出了沃尔德检验结果。

表 8.7 基于设计与未加权 SRS 分析选择下，模型 2 中交互变量的显著性沃尔德检验(X^2(b))

变 量	df	基于设计		未加权 SRS	
		X^2_{des}	p-值	X^2_{bin}	p-值
SEX * PHYS	1	2.39	0.121 8	3.97	0.046 3

当使用忽略整群效应的未加权 SRS 分析选择时，SEX 与 PHYS 的交互效应是显著的，因而得到了比基于设计的分析选择下更复杂的模型。这些结果显示，即使整群效应并不严重，就如这里仅仅是中等规模的组群设计效应，忽略它是令人警醒的。

让我们转向表 8.2 中比例的线性模型所对应的基于设计的分析。在这种分析选择下，ANOVA 模型的对数与线性类型得到的结果非常相近，因为比例的波动离 0.5 不远。选择了主效应模型(模型 1)，模型拟合、残差以及模型项的显著性与对数模型的很相似。但是，模型系数估计值不同，且其解释也不同。对于偏参数化的对数模型，估算系数表示相应组别在比率对数的刻度上偏离作为参照组群的拟合比率对数的截距的差异。对于线性模型，估算系数表示相应组别在线性刻度上偏离估算作为参照组群的拟合比例的截距的差异。

因此，线性模型涉及对模型系数估计值更直截了当的解释。在模型 1 中，这些估计值如下：

$\hat{b}_1 = 0.5705$ （截距）

$\hat{b}_2 = -0.1172$ （SEX = 男性的差异效应）

$\hat{b}_3 = -0.0355$ （AGE = 44 以下的差异效应）

$\hat{b}_4 = 0.0650$ （PHYS = 1 的差异效应）

因此，工作环境风险较小的老龄女性，落入精神压力较高组的拟合比例为 0.57，而相应的年龄组男性的比例为 0.57 − 0.12 = 0.45。最大拟合比例为 0.57 + 0.07 = 0.64，对应工作环境风险较大的老龄女性。同时，这些拟合比例与相应的对数模型中的结果很接近。

8.4 对数与线性回归

在与 GWLS 方法相似的分析选择下,类似然(PML)方法通常用于复杂调查数据的对数分析。但是,PML 方法的适用范围更广,不仅包括二分或是多值回应变量的组群比例模型,还包括以连续变量为预测变量的常见的回归类型的情形。在本小节中,我们首先讨论组群比例的 PML 分析,然后是二分回应变量加上混合连续与类别预测变量的更通用的对数模型。最后,在 ANCOVA 情形中,给出一个连续回应变量的线性模型例子。

在模型系数的 PML 估算及其渐进协方差矩阵估算中,我们使用改进的极大似然方法(ML)。在简单随机样本的 ML 估算中,我们的对象是未加权的观测个案。根据标准分布假设,可以构造合适的似然方程,以获取模型系数的 ML 估计值及其协方差矩阵估计值。根据这些估计值,可以使用标准似然比率(LR)与基于二项的沃尔德检验统计量,来检验模型是否满意以及模型系数的线性假设。

在更复杂的、涉及元素加权与整群的设计中,模型系数的 ML 估计值与相应的协方差矩阵估计值并不一致,并且,标准检验统计量并不满足合适自由度下的渐进卡方分布。为了得到模型系数的一致估算,改进标准似然方程以涵盖加权的观测个案。除此之外,在剔除整群效应后来构建 PML 估计值一致的协方差矩阵估计值。使用这些一致的估计值,可以推导出合适的渐进卡方检验统计量。

在与假定一个二分回应变量与一组类别变量的 GWLS 方法相似的情形下,可以很简便地介绍 PML 方法。数据安排成一个表 8.1 所示的多维表格,含有 u 个组群。我们的目标是用模型来解释各个组群的组群比例估计值 $\hat{p}_j$ 的协变。模型的类型是式 8.1 与式 8.2 给出的对数模型。组群比例的 PML 对数模型,包括含有类别预测变量的对数 ANOVA,ANCOVA 及回归模型。使用相应的组群比例估计向量及其协方差矩阵估计值,可以在如前介绍的任何分析选择下估算它们。其建模过程与 GWLS 方法的过程相同。基于设计的分析选择,为复杂调查给出了通常有效的 PML 对数分析。在实际中,调查分析的专业软件提供基于设计的 PML 对数分析。

基于设计的二项类似然(PML)方法

在基于设计与加权 SRS 分析选择下,对数模型 $F(\mathbf{p}) = \mathbf{Xb}$ 的 s 个的模型系数 b_k 的向量 $\mathbf{b}$ 的一致类似然(PML)估计值 $\hat{\mathbf{b}}_{pml}$,是通过叠代计算如下 PML 估算方程得到的,

$$\mathbf{X'Wf}(\hat{\mathbf{b}}_{pml}) = \mathbf{X'W\hat{p}}, \quad (8.24)$$

其中，$\mathbf{W}$ 是 $u \times u$ 对角权重矩阵，权重 $w_j = \hat{n}_j$ 在主对角线上；$\mathbf{f} = \exp(\mathbf{Xb})/[1 + \exp(\mathbf{Xb})]$ 是比率对数函数的倒函数。在式 8.24 中，重要的是使用加权组群样本规模 $\hat{n}_j$ 与加权比例估计值 $\hat{p}_j$，而非未加权 SRS 分析选择下 ML 方法中的未加权的 n_j 与 $\hat{p}_j^{\,U}$。这是为了 PML 估算公式中的一致性。相应的 GWLS 估计值向量（式 8.5）可以用作 PML 叠代的起始值。注意，在 ANOVA 模型的线性表达中，函数向量 $\mathbf{f}(\hat{\mathbf{b}}_{pml})$ 在$\hat{b}_k$ 中可以是线性的，因而并不需要叠代计算。此后在本小节中，为求简洁，我们用$\hat{\mathbf{b}}$来表示对数模型系数的 PML 估计值向量。

在基于设计的分析选择与在加权 SRS 分析选择下，由于 PML 估计值向量$\hat{\mathbf{b}}$相等，所以拟合比率对数 $\hat{\mathbf{F}} = \mathbf{X}\hat{\mathbf{b}}$与拟合比例 $\hat{\mathbf{f}} = F^{-1}(\mathbf{X}\hat{\mathbf{b}})$都相等。估算的比率比可以在模型的偏参数化中，以 $\exp(\hat{b}_k)$ 得到，它在两个分析选择下也相等。在两个分析选择下，拟合比例$\hat{f}_j = \hat{f}_j(\hat{\mathbf{b}})$的估算使用以下公式，

$$\hat{\mathbf{f}} = \mathbf{f}(\hat{\mathbf{b}}) = \exp(\mathbf{X}\hat{\mathbf{b}})/(1 + \exp(\mathbf{X}\hat{\mathbf{b}}))。 \quad (8.25)$$

让我们推导，在加权 SRS 与基于设计分析选择下，由式 8.24 计算的 PML 估计向量$\hat{\mathbf{b}}$的 $s \times s$ 协方差矩阵估算公式。假定简单随机抽样，这一协方差矩阵估算公式为，

$$\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}}) = (\mathbf{X'W\hat{\Delta}WX})^{-1}, \quad (8.26)$$

其中，对角 $u \times u$ 矩阵 $\hat{\Delta}$ 的对角元素是二项类型方差 $\hat{f}_j(1-\hat{f}_j)/\hat{n}_j$。式 8.26 中的二项协方差矩阵估计值对于涉及整群的复杂抽样设计并不一致。对于这样的设计，我们推导出在基于设计分析选择下有效的、更加复杂一致的协方差矩阵估算公式，

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}}) = \hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}})\mathbf{X'W}\hat{\mathbf{V}}_{des}\mathbf{WX}\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}})。 \quad (8.27)$$

这一估算公式有着“夹逼”的形式，而比例向量 $\hat{\mathbf{p}}$ 基于设计的协方差矩阵估计值 $\hat{\mathbf{V}}_{des}$则是“夹心形式”。

与 GWLS 方法一样，使用相应的 PML 估计值$\hat{b}_k$ 的方差估计值 $\hat{v}_{des}(\hat{b}_k)$与 $\hat{v}_{bin}(\hat{b}_k)$，根据式 8.7 可以计算，在基于设计与加权 SRS 分析选择下的，比率比估计值 $\exp(b_k)$的近似置信区间。同样的，与 GWLS 方法一样，根据式 8.23 可以得到模型系数$\hat{b}_k$ 的设计效应估计值 $\hat{d}(\hat{b}_k)$。

基于设计分析选择下的拟合比率对数 $\hat{\mathbf{F}}$ 与拟合比例 $\hat{\mathbf{f}}$ 的一致性协方差矩阵估计值 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{F}})$与 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{f}})$的表达式，与等式 8.8 与式 8.9 给出的 GWLS 方法

下的表达式相似。当然,等式中必须使用式 8.27 中的 PML 类似的 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ 与相应的矩阵 $\hat{\mathbf{H}}$。在加权 SRS 分析选择下,在等式中使用二项估算公式 8.26 来代替基于设计的项,可以相似地推导出协方差矩阵估计值 $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{F}})$ 与 $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{f}})$。

在基于设计分析选择下,合适的残差分析需要残差协方差估计值。这一 $u \times u$ 的估计值为,

$$\hat{\mathbf{V}}_{res} = \mathbf{A}\hat{\mathbf{V}}_{des}\mathbf{A}', \qquad (8.28)$$

其中,矩阵 $\mathbf{A}$ 由以下公式获得,

$$\mathbf{A} = \mathbf{I} - \hat{\Delta}\mathbf{WX}(\mathbf{X}'\mathbf{W}\hat{\Delta}\mathbf{WX})^{-1}\mathbf{X}'\mathbf{W}$$

其中的 $\mathbf{I}$ 是一个 $u \times u$ 的单位矩阵。使用这一估计值,可以计算式 8.22 中的基于设计的标准残差。

因此,PML 公式与 GWLS 方法下的公式有很多相似点。主要的差别是模型系数的估计值及其协方差矩阵估计值的计算方式。在检验程序中的相似性更明显。所有在 GWLS 方法下推导出的检验统计量,也适用于 PML 方法。

在基于设计的分析选择下,可以使用式 8.11 给出的基于设计的沃尔德统计量 X^2_{des} 来检验模型拟合度。当更仔细地检验模型拟合度时,可使用沃尔德统计量 $X^2_{des}(overall)$ 与 $X^2_{des}(gof)$ 的 PML 类似统计量。用于模型参数的线性假设的沃尔德统计量(式 8.13 与式 8.14)同样适用。最后,在不稳定的情形下,可以使用 F-校正的沃尔德与拉奥-斯科特统计量(式 8.16 到式 8.20)。应当注意到,在基于设计的分析选择下,计算这些统计量时,必须使用式 8.24 中的 PML 估计值,以及相应的式 8.27 中的协方差矩阵估计值。常用的复杂调查数据对数分析软件含有这些检验统计量。

在加权与未加权 SRS 分析选择的检验过程中,检验统计量中的基于设计的协方差矩阵估计值由相应的二项协方差矩阵估计值所替换。作为沃尔德统计量的替代,可以使用 LR 检验统计量。它在基于设计的分析选择下,应当做拉奥-斯科特修正。与式 8.14 的基于二项的沃尔德统计量相似,LR 检验统计量的二阶修正是渐进卡方分布的检验统计量。式 8.28 中的残差协方差矩阵估计值,可用来推导修正值的通用设计效应矩阵估计值。

复杂调查的 PML 方法的主要适用范围是基于设计的分析选择。在检查加权与群内相关对于模型系数标准误估计值以及沃尔德检验统计量的 p-值的影响作用时,加权与未加权 SRS 分析选择被用作参照。

对数回归

PML 方法也可以用于复杂调查中二分回应变量的严格回归类型的对数分析。这里的预测变量为连续性的。在对数回归中,我们的对象是元素层次的数据,并没有加总成一个多维表格。所以,连续预测变量的取值构成了对数

回归模型的 $n \times s$ 模型矩阵 $\mathbf{X}$ 的列。但是,所有其他 PML 估算的元素均不变。得到一致的 PML 估计值及其一致的协方差矩阵估计值的方法与基于设计的分析选择下所描述的相似。同时,纳入类别变量到对数回归模型中,可以进行对数 ANCOVA 分析。而连续与类别预测变量的交互效应项也可以包括到其中。

对数回归的建模过程是,使用主题标准或是潜在显著性来逐步将预测变量引入到模型中。在这一过程中,与前面的一样,可以使用模型系数的 t-检验 $t_{des}(b_k)$ 或是相应的沃尔德检验 $X^2_{des}(b_k)$。而在基于设计分析选择下,这些检验统计量的渐进特征没有变化。

在样本整群数目较小的小样本情形下,式 8.27 中的估计值 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ 的不稳定性可能会破坏模型系数的检验统计量的分布特征。这时,可以使用通常的沃尔德检验统计量的自由度与 F-校正。

通用估算方程(GEE)方法也可以用于复杂调查数据的对数模型。这一方法使用多变量准似然技术来估算模型系数,而将群内相关当成干扰。使用估算的群内相关结构,可以得到模型系数的协方差矩阵的"抗扰"估计值。它基本上与 PML 方法中的"夹逼"形式类似。因此,GEE 方法可用来排除整群效应。我们将简要描述这一方法,并给出一个 OHC 调查的对数 ANCOVA 例子。

GEE 方法最早是在历时研究中拟合通用线性模型时,用来剔除可能的观察个案间的相关(Liang and Zeger,梁与齐格尔,1986)。梁等(Liang et al.,1992)与迪格尔等(Diggle et al.,2002)进一步描述与演示了这一方法。

我们给出两种 GEE 方法。初级 GEE 方法假定独立相关,并与标准的 PML 方法相连。在估算回归系数时,假定群内观察个案独立;在估算回归系数估计值的协方差矩阵时,允许群内相关。在协方差估算中,使用了"夹逼"形式的估算公式。在更高级的 GEE 方法中,假定了可交换的相关结构,在回归系数的估算与回归系数估计值的协方差矩阵的估算中,都允许群内相关。在这里,估算了一个"工作"群内相关,并将它纳入回归系数与回归系数估计值的协方差矩阵的估算中。

通用线性模型可以简洁地写成

$$\mathrm{E}_{\mathrm{M}}[g(\mathbf{y})] = \mathbf{Xb}, \tag{8.29}$$

其中,E_{M} 是模型的期望值,而函数 g 是所谓的表示回应变量向量 $\mathbf{y}$ 与模型线性部分 $\mathbf{Xb}$ 之间关系的链接函数。链接函数的特例有,分别用于连续回应变量的线性模型、二分回应变量的对数模型以及计数数据的对数线性模型的单位函数、逻辑斯蒂与对数函数。

群内观察个案的协方差结构为,

$$\mathbf{V}_i = \phi \mathbf{A}_i^{1/2}\mathbf{R}(\alpha)\mathbf{A}_i^{1/2}, \quad i = 1, \cdots, m, \tag{8.30}$$

其中，$\mathbf{A}_i$ 是整群 i 中方差 $V(y_k)$ 的对数矩阵，$\mathbf{R}(\alpha)$ 是整群 i 中观察个案的、由（可能为向量的）相关参数 α 设定的“工作”相关矩阵。参数 ϕ 表示相应指数族分布函数成员的离散参数。在独立相关假设下，“工作”相关矩阵的所有对角线之外的元素 α 为 0。在群内成对观察个案可交换相关的假设下，参数 α 为一个换算因子，并需要估算。在得到估计值 $\hat{\mathbf{b}}$ 的过程中，通常使用牛顿-拉夫逊类型的算法。通过使用“夹逼”类型的估算公式（参见等式 8.27）得到协方差矩阵 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$。元素权重可以纳入到 GEE 估算过程中。GEE 及类似的加权方法可以应用于合适的复杂调查分析的软件中。

我们演示了 GEE 方法生成模型参数及其协方差矩阵的一致估计值。这是独立于“工作”相关结构的正确设定。在接下来的两个例子中，我们首先将 PML 方法应用到对数 ANCOVA；然后将假定可交换群内相关结构的 GEE 方法应用于对数 ANCOVA。对于 OHC 调查数据对数模型的 PML 与 GEE 方法进一步训练，读者可参考本书的扩展网页。

范例 8.2

对数 ANCOVA 的 PML 方法。在范例 8.1 中，GWLS 方法被用来拟合多维表格中比例的对数 ANOVA 模型。让我们考虑比它略微普通的分析选择。将一些预测变量当成连续变量引入模型，我们现在用 PML 方法拟合对数 ANCOVA 模型。给定有效的 PML 分析，我们使用基于设计的分析选择。

二分回应变量 PSYCH 测量高精神压力，变量 AGE，PHYS（物理工作环境）与 CHRON（慢性病）被当成连续预测变量。AGE 的测量刻度是岁数，PHYS 与 CHRON 则为二分变量。共有 4 个预测变量，其中的 SEX 被当成一个定性预测变量。因此，也可以检查 SEX 与 AGE，PHYS，CHRON 之间的交互效应。

含有 SEX，AGE，PHYS 与 CHRON 主效应以及 SEX 与 AGE 交互效应项的模型，被当成最后模型。这是因为其他交互效应项在 5% 水平并不显著。表 8.8给出了模型系数的结果。

使用与范例 8.1ANOVA 模型相似的估算的系数 $\hat{b}_k$ 以及相应的模型矩阵 $\mathbf{X}$，拟合 ANCOVA 模型可以写成：

$$F(\hat{f}_l) = \hat{b}_1 + \hat{b}_2(\mathrm{SEX})_l + \hat{b}_3(\mathrm{AGE})_l + \hat{b}_4(\mathrm{PHYS})_l + \hat{b}_5(\mathrm{CHRON})_l + \hat{b}_6(\mathrm{SEX} * \mathrm{AGE})_l,$$

其中，$l = 1, \cdots, 7\,841$，$F(\hat{f}_l) = \log[\hat{f}_l/(1-\hat{f}_l)]$。从 7 841 × 6 的模型矩阵 $\mathbf{X}$ 相应的列中，可以得到模型项的数值。这里，SEX，PHYS 与 CHRON 为二分，AGE 有原来的数值（岁数）。注意与 ANOVA 模型矩阵相比，ANCOVA 模型矩阵的差异。

模型系数的 t-检验显示，所关注的预测变量——物理工作环境与慢性病——的系数都强烈地与感受精神压力相关。高工作风险与患慢性病的人，比健康且低工作风险的人更容易感受精神压力。注意，控制性别与年龄的 CHRON 的系数$\hat{b}_5$ 比 PHYS 系数$\hat{b}_4$ 更大。因此，在模型中，慢性病是精神压力的一个更重要的预测变量。从表 8.8 中给出的比率比(OR)，也可以看出这一点。

表 8.8 整体心理压力基于设计的对数 ANCOVA(PML 方法)

							比率比的95%置信区间	
模型变量	β 系数	设计效应	标准误	t-检验	p-值	比率比	下限	上限
截距	0.196 4	1.56	0.157 2	1.25	0.212 7	1.22	0.89	1.66
性别								
男	-0.992 6	1.43	0.203 3	-4.88	0.000 0	0.37	0.25	0.55
女*	0	n.a.	0	n.a.	n.a.	1	1	1
年龄	-0.004 6	1.55	0.004 1	-1.12	0.262 4	1.00	0.99	1.00
身体健康风险	0.276 5	1.39	0.059 6	4.64	0.000 0	1.32	1.17	1.48
慢性病	0.564 1	1.17	0.057 5	9.82	0.000 0	1.76	1.57	1.97
性别,年龄								
男	0.013 1	1.41	0.005 1	2.56	0.011 1	1.01	1.00	1.02
女*	0	n.a.	0	n.a.	n.a.	1	1	1

*:参照组;参数设定为0。

n.a.:不适用。

比率比及其近似95%的置信区间(括弧内)为,

PHYS: 比率比 = exp(0.276 5) = 1.32 (1.17, 1.48),

CHRON: 比率比 = exp(0.564 1) = 1.76 (1.57, 1.97)。

我们可以得出结论，控制性别、年龄与慢性病，高工作风险的人感受精神压力的概率为低工作风险的人的 1.3 倍。这一结论与范例 8.1 相似。在那里，获得了高度相关的比率比与置信区间。同时，控制性别、年龄与工作风险，患有慢性病的人感受精神压力的概率为健康人的 1.8 倍。由于两者的 95% 置信区间均不包含数值 1，相应的比率比显著地(在 5% 水平)与 1 不同。应当注意到，基于二项的预测变量 PHYS 的置信区间将要狭窄一些，它的设计效应估计值比 CHRON 的要大。

SRS 选项下得到与基于设计分析相同的最终模型。但是，检验统计量的观测值要大些。因此，得到更大的检验结果。

最后，让我们在当前模型中进一步检查精神压力更大组别的拟合比例$\hat{f}_l$。通过将比例根据模型中的预测变量制图，图8.2给出了结果的小结。男性的拟合比例随年龄增加而增加，女性的则随年龄增加而降低。给定年龄，患有慢性

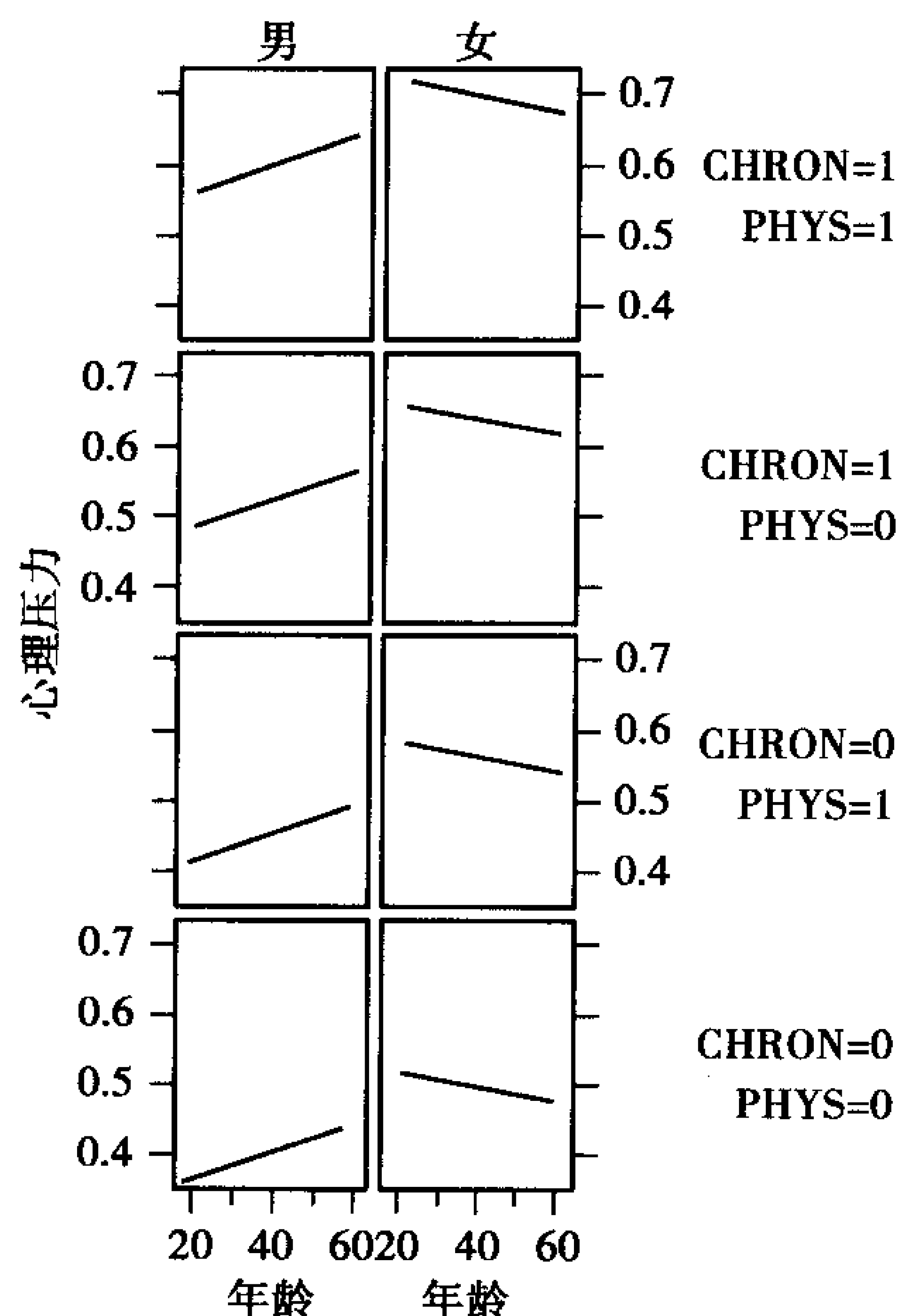


图 8.2 最终对数 ANCOVA 模型中落入高心理压力组的拟合比例

病的以及工作风险高的比例，高于参照组。同时，在所有相应的组群中，女性的拟合比例倾向于高于男性的拟合比例。但随年龄的增加，差异有所降低。

范例 8.3

对数 ANCOVA 的 GEE 方法。让我们进一步考虑范例8.2中的分析选择。其中的对数 ANCOVA 模型用 PML 方法来拟合。使用假定群内成对观测个案间可交换相关的 GEE 方法，我们现在来拟合对数 ANCOVA 模型。与范例 8.2 相似，我们的回应变量为测量精神压力的二分 PSYCH。变量 SEX 被当成类别预测变量包括在模型中。AGE，PHYS（物理工作环境）与 CHRON（慢性病）被当成连续预测变量。AGE 的测量刻度是岁数，PHYS 与 CHRON 则为二分变量。我们拟合范例8.2中相同的模型。

在表 8.9 中，与范例 8.2 中 PML 方法的对数 ANCOVA 的比较显示，结果非常相似，我们的推论结论也相同。当然也有不同。第一，估算的 beta 系数改变了。除了 CHRON 效应，估计值的绝对值大于 PML 方法的结果。标准误也小于 PML 方法的结果。所以，t-检验观测值倾向于更大，得到的检验比 PML 方法也更大。这些不同是由于假定了可交换的相关结构的 GEE 方法中，观察个案之间的相关影响了 beta 参数的估算。

“工作”群内相关的估计值为 $\hat{\alpha} = 0.0189$。使用表达式 $\text{deff} = 1 + (\overline{m} - 1)\hat{\alpha}$，得到平均设计效应为 1.57，其中的 $\overline{m}$ 是平均整群规模。

表 8.9　整体心理压力基于设计的对数 ANCOVA(可互换组内相关系数结构的 GEE 方法)

模型变量	β 系数	设计效应	标准误	t-检验	p-值
截距	0.229 2	1.44	0.152 4	1.50	0.133 8
性别					
男	-1.029 0	1.36	0.200 0	-5.14	0.000 0
女*	0	n. a.	0	n. a.	n. a.
年龄	-0.005 7	1.43	0.003 9	-1.45	0.148 9
身体健康风险	0.301 1	1.31	0.058 7	5.13	0.000 0
慢性病	0.556 9	1.14	0.056 8	9.81	0.000 0
性别,年龄					
男	0.014 4	1.33	0.005 0	2.88	0.004 4
女	0	n. a.	0	n. a.	n. a.

*:参照组;参数设定为 0。

n. a.:不适用。

连续回应变量的线性模型

我们广泛地讨论了复杂调查中二分回应变量的建模。使用了 GWLS, PML 与 GEE 方法,涵盖了类别变量的对数与线性模型以及连续预测变量的对数模型。这样的多变量模型最常见于社会与卫生学科的分析性调查中。但在一些情形下,需要用模型来解释定量或是连续的回应变量,比如医生访问次数或是血压。我们简要讨论这些情形下的多变量分析特征,并给出一个线性 ANCOVA 特例的演示例子。

线性模型为一个连续回应变量与一组预测变量的分析情形给出了简便的分析方法。在范例 8.2 与 8.3 的这样的分析选择中,对数 ANCOVA 模型分析了二分的 PSYCH。原有的精神压力的连续变量也可以被当成回应变量,而需要使用线性 ANCOVA 模型。对于简单随机样本,分析是基于普通最小二乘法(OLS)估算的标准线性建模。对于基于整群抽样的 OHC 调查数据,基于设计的加权最小二乘法(WLS)估算能够给出合适的线性模型。

基于设计的分析选择下,线性模型面临与前面建模技术相似的复杂性。与已经介绍过的 GWLS,PML 及 GEE 建模方法相比,在估算技术与检验过程中,没有新的元素。所以,我们的首要目标是模型系数的一致性估算以及估算的系数的协方差矩阵的一致性估算。这些需要用合适的元素权重来加权,而模型系数估计值的协方差矩阵的估算需要剔除整群效应的影响。

线性回归模型可以简洁地写成矩阵模型形式如下,

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \tag{8.31}$$

其中,$\mathbf{y}$ 是回应变量取值向量,$\mathbf{X}$ 是模型矩阵,$\mathbf{b}$ 是将要估算的回归系数向量,而 $\mathbf{e}$ 是随机误差向量。

在基于设计与加权 SRS 分析选择下，向量 $\mathbf{b}$ 的一致估算是对以下加权正态方程求解，

$$\mathbf{X'WX}\hat{\mathbf{b}} = \mathbf{X'Wy}, \tag{8.32}$$

其中，$\mathbf{W}$ 的对角元素是换算后的元素权重 w_l^{**}。在未加权的 SRS 分析选择下，权重均为 1，估算简化成通常的 OLS 估算。$\hat{\mathbf{b}}$的 WLS 估算公式为，

$$\hat{\mathbf{b}} = (\mathbf{X'WX})^{-1}\mathbf{X'Wy}。 \tag{8.33}$$

正如基于设计的比例的 PML 方法，在基于设计的分析选择下，使用“夹逼”类型估算公式来一致估计估计值$\hat{\mathbf{b}}$的协方差矩阵。与用于比例的对数与线性模型的 GWLS，PML 及 GEE 方法中的沃尔德与 F-统计量相似，使用这样的检验统计量，可以进行模型拟合度与模型系数的线性假设的检验。

在实际中，合适的调查分析的计算机软件，可以简便地进行基于设计的线性建模。

范例 8.4

感受到的精神压力的 WLS 线性建模方法。在范例 8.2 与 8.3 中，对精神压力二分变量 PSYCH 拟合了对数 ANCOVA 模型。现在，对原有的变量 PSYCH 拟合一个线性 ANCOVA 模型。原有的 PSYCH 是 9 个精神症状的第一个标准化主成分赋值。因此，PSYCH 的均值为 0，方差为 1。但是，PSYCH 的分布有些偏态：数据中的许多人没有经历任何问题中的精神症状。PSYCH 的取值区间为(-1, 4.7)，分布的中位值为 -0.4。

在线性 ANCOVA 模型中，我们引入与前面两个例子相同的变量作为潜在的预测变量。预测变量 SEX 被当成是定类变量，而 AGE，PHYS 与 CHRON 被当成是连续的。我们也考察 SEX 与连续预测变量间成对的交互效应。用 WLS 方法来拟合模型，建模过程生成了与范例 8.2 与 8.3 相似的 ANCOVA 模型，即，所有主效应和 SEX 与 AGE 的交互效应显著。

与二分 PSYCH 的对数模型一样，使用估算的系数$\hat{b}_k$ 以及相应的模型矩阵 $\mathbf{X}$，PSYCH 的拟合的线性 ANCOCA 模型可以写成：

$$\hat{f}_l = \hat{b}_1 + \hat{b}_2(\mathrm{SEX})_l + \hat{b}_3(\mathrm{AGE})_l + \hat{b}_4(\mathrm{PHYS})_l + \hat{b}_5(\mathrm{CHRON})_l + \hat{b}_6(\mathrm{SEX} * \mathrm{AGE})_l,$$

其中，$l = 1, \cdots, 7\ 841$，从范例 8.2 中的模型矩阵 $\mathbf{X}$ 得出所有模型项的数值。表 8.10 给出了以连续变量精神压力为回应变量的 ANCOVA 模型系数的结果。

模型系数的符号和 t-检验的结果与范例 8.2 中相应的比率对数 ANCOVA 的结果相似。但是，比率对数模型的系数的解释有所不同。在比率对数 ANCOVA中，我们的对象是二分回应变量的比率刻度；现在，我们的对象是线性刻度上的连续变量。因此，线性 ANCOVA 模型的系数可以作通常意义上的

线性回归的解释。

表 8.10 整体心理压力基于设计的线性 ANCOVA(WLS 方法)

模型变量	β 系数	设计效应	标准误	t-检验	p-值
截距	-0.012 1	1.70	0.083 1	-0.15	0.884 6
性别					
男	-0.497 5	1.48	0.099 7	-4.99	0.000 0
女*	0	*n. a.*	0	*n. a.*	*n. a.*
年龄	-0.000 1	1.60	0.002 1	-1.02	0.980 4
身体健康风险	0.177 2	1.37	0.029 0	6.11	0.000 0
慢性病	0.392 2	1.17	0.029 4	13.33	0.000 0
性别,年龄					
男	0.005 7	1.39	0.002 5	2.25	0.025 2
女	0	*n. a.*	0	*n. a.*	*n. a.*

*:参照组;参数设定为 0。

n. a.:不适用。

在加权 SRS 分析选择下,将得到相同的 ANCOVA 模型,而模型系数的结果也相同。但是,模型系数的标准误要小些。这是因为,设计效应估计值 $\hat{d}(\hat{b}_k)$ 大于 1。但是,这并不影响 t-检验结果的推论。

虽然回应变量的分布有些偏态,但由于它的连续性,变量 PSYCH 为演示线性建模提供了一个很好的机会。在实际中,经常遇到计数变量——例如,在某一时间段访问医生的次数;或是其他相似的分布很偏的变量。在拟合线性模型之前,对这样的变量的建模可能涉及诸如计量经济学常用的对数或是博克斯-考克斯转换的对称化转换。另外,线性模型的表达式可能并不适合于这样的变量。所以,需要使用其他回归建模技术,如,解释超泊松协变的泊松回归与负二项分布模型。这些方法属于相关回应变量的通用线性模型。对于这些模型,从干扰的角度,可以成功地使用类似然与通用估算方程的方法。

分解方法

到目前为止,讨论的多变量分析方法均是干扰或是聚合方法。其目标是从分析结果中剔除整群效应的干扰,以便得到一致的估算与有效的检验结果。另一方面,在分解方法中,群内相关结构本身就很有意义,对这些相关系数的估算构成分析的重要组成部分。这常见于对象为垂直分级结构数据的社会与教育调查中。村庄、公司或是学校的整群构成了常见的分级结构的例子。

有高级的方法来做垂直分级结构数据中的群内相关的回应变量的多变量分析。多层级建模(multi-level modelling)是建立在通用线性混合模型的基础之上的,它将某些随机效应纳入了模型之中。这是本书没有讨论的新的一族模型。前面所有的模型中,模型参数均被当成是固定效应。多层级模型的应

用主要用在教育调查的连续回应变量的线性建模中,其中的学校与班级被当成整群(Goldstein, 1987,2002)。对二分或多级回应变量的多级模型也得到了发展,也有了合适的算法。在章节 9.4 中,我们在分整群教育数据中,使用多层级模型与一个连续回应变量。在那里,将给出这种方法的简要介绍。

8.5 本章小结与更多的文献

小 结

在本章里,主要从干扰方法的角度,讨论了群内相关回应变量的线性与比率对数模型。其主要的目的是从估算与检验结果中剔除群内相关的影响。但是,这些效应的严重性,随着不同抽样设计而变化。因而,介绍了各种合适的实际分析中的分析选择。

基于设计的分析选择,给出了一个复杂调查的多变量分析在总体上有效的分析选择。在这种分析选择下,抽样实际的复杂性——包括整群、分层与加权,被合适地化解了。基于设计的分析选择下的分析,需要元素层次的数据及合适的调查分析软件。同时,对于分层元素抽样与简单随机抽样,加权与未加权 SRS 分析选择也可以做有效的分析。在加权 SRS 分析选择下,仅包括了加权;而在未加权 SRS 分析选择下,则忽略了所有抽样复杂性。因此,这些分析选择并不适合于整群设计的复杂调查。

在任意分析选择下,二分或是多值回应变量的比率对数与线性 ANOVA, ANCOVA 及回归分析,都可以对安排成多维表格的数据作通用加权最小二乘法(GWLS)估算。在基于设计的分析选择下使用的 GWLS 方法,对复杂调查中这样的表格给出了有效的分析。为了得到可靠结果,需要元素与整群样本数目够大。这样的要求在大规模分析性调查中——诸如,基于分层整群抽样设计的 OHC 调查——得到满足。如果样本整群数目较小,可能产生不稳定的问题,使得估算与检验结果并不可靠。对检验统计量使用合适的修正,可以成功地解决这样的问题。

类似然估算方法(PML)可用于与 GWLS 相似的分析选择。但它主要用于 GWLS 并不合适的连续预测变量的对数回归中。在基于设计的分析选择下,PML 对复杂调查给出了有效的比率对数分析。样本整群数目较大对于 PML 方法也是有益的。与 GWLS 方法一样,对于 PML,也有相似的修正不稳定问题的方法。在分析 OHC 调查的二分回应变量中,我们使用 PML 方法于对数 ANCOVA 模型。

PML 方法不仅仅涵盖了对数回归模型,还包括通用线性模型家族的其他

类型的模型。因此,也涵盖了连续回应变量的线性模型。我们简要地介绍了通用估算方程方法(GEE)。我们将假定了整群内可交换的相关结构的 GEE 方法用于二分回应变量的对数 ANCOVA 模型,得到了与 PML 方法相似的结果。

对 OHC 调查某些多变量分析选择的个案研究表明,化解抽样复杂性——特别是整群效应,对于有效推论十分关键。在第 9 章中,我们将演示这一重要结论,并给出其他复杂调查数据的个案分析。

对于群内相关的回应变量,干扰或是分解方法在不同类型的多变量分析选择下,给出了合理与可控的分析策略。在另外的分解方法中,与模型系数一样,群内相关被当成需要估算的、本身有意义的参数。我们简要讨论了适用于垂直分级结构数据的多层级模型。在下一章,我们将演示多层级模型的方法。

更多的文献

复杂调查的多变量分析在文献中得到了相当的关注。可以从宾德(Binder,1983)、拉奥与斯科特(Rao and Scott,1984,1987)、罗伯茨等(Roberts et al.,1987)、拉奥等(Rao et al.,1989)及斯科特等(Scott et al.,1990)相关论述中看出方法论上的推进。他们讲解了复杂调查中,类别变量的比率对数与相关分析的加权最小二乘法、类似然与准似然方法。斯金纳等(Skinner et al.,1989)的编著讨论了聚合与分解方法在多变量分析中的许多进展。拉奥与托马斯(Rao and Thomas,1988)及科恩与格劳巴德(Korn and Graubard,1999)给出了这些方法的应用。钱伯斯与斯金纳(Chambers and Skinner,2003)的编著包括了几篇对复杂调查数据分析方法持不同观点的文章。

拉奥等(Rao et al.,1993)讨论了二级整群样本的回归分析。宾德(Binder,1992)讲解了复杂数据比例概率模型的拟合。宾德(Binder,1991)讨论了无应答类别变量的分析,而格林等(Glynn et al.,1993)讨论了线性模型的多元推断。

多级模型在戈尔茨坦(Goldstein,1987,1991)中得以介绍,并在戈尔茨坦与拉斯巴克(Goldstein and Rasbash,1992)及戈尔茨坦(Goldstein,2002)中得到了发展。普费弗曼等(Pfeffermann et el.,1998)讨论了多级的加权问题。通用估算方程的建模由梁与齐格尔(Liang and Zeger,1986)引入,并在梁等(Liang et al.,1992)及迪格尔等(Diggle et al.,2002)中得到进一步发展。霍顿与利普斯茨(Horton and Lipstz,1999)讨论了软件,齐格勒等(Ziegler et al.,1998)回顾了 GEE 方法的文献。布雷斯洛与克莱顿(Breslow and Clayton,1993)给出了通用线性混合模型框架中近似推论的一般结果。克莱顿等(Clayton et al.,1998)及费德等(Feder et al.,2000)讨论了复杂历时调查数据的分析。

第9章

更多详细的例子

More Detailed Case Studies

我们选择4个例子来主题明确地演示本书中讨论的调查方法。第一个例子(章节9.1)讨论一个长时期调查数据收集中的质量监督。本书前面介绍了若干统计量作为质量标志。乘客交通调查提供经验结果。其资料收集历时一整个年份,每月的样本规模相等。

第二个例子(章节9.2)来自于一个商务调查,是一个常见于商业统计中的解决抽样框问题的例子。在两个不同的框架中,讨论某些职业组的平均年薪的估算。这将导致混合类型数据收集的策略,3个季度的数据来源于普查,1个季度的数据来源于抽样调查。另外,我们对于商业调查的分析表明,在计算样本中雇员层次的统计量而抽样单位是公司时,应当剔除整群影响效应。

在社会经济调查中(章节9.3),对一个来源于以家庭户为整群的整群抽样设计类别数据,拟合比率对数模型。强调的重点,并不仅仅在于指出剔除整群效应的重要性,还在于选择满意的分析模型的类型的重要性。这里使用了方差分析与回归类型的对数模型,它们得出了不同的结论。

在最后一个例子中(章节9.4),利用从多国教育调查的整群调查数据,使用一个多层级回归模型来介绍并演示垂直分级结构数据的模型分析方法。这些模型与前面例子中使用的干扰方法不同。在多层级建模中,总体的分级结构反映到了模型结构中。结果中也包括有趣的国家间的比较。

9.1 长期交通调查中的质量监督

数据收集在很多调查中历时一个较长时期,比如,一整年。这种社会调查中,较好的例子是整个样本等分为12份的消费者意愿调查与旅行或是迁移调查。这种调查策略有两个目的:逐月收集截面数据,以及组合成年度数据以期发现这些现象的季节性、周期性或是趋势性特征。在这种调查中,一个主要问题是,在整个调查时期内保持一致的数据质量。因此,数据收集过程中的质量

监督变得很重要。

在这个例子中，使用了一组 20 个统计质量标识变量来监控每个数据收集周期中可能的偏差。这些标识变量涵盖了抽样与非抽样误差的重要方面。在前面已经定义一些标识变量。比如，离异系数、涵盖率、应答率以及组内相关。更多的不同的调查误差可以参见格罗夫斯（Groves，1989）。考克斯等（Cox et al.，1995）讨论了商业调查中的这一问题。比默与利伯格（Biemer and Lyberg，2003）给出了一个调查质量的非技术性介绍。

乘客交通调查

我们用一个历时较长的调查——由芬兰交通与传播部在 1998—1999 年进行的乘客交通调查，来演示质量标识变量的使用。这一调查总计有 18 250 个抽样单位，等分为 12 个 1 500 的月度份额。数据由计算机辅助的电话访谈（CATI）来收集。帕斯蒂能（Pastinen，1999）报告了主要结果与调查过程。

为了监督质量的一致性，在月度数据份额收集中，有两个报告表格。每个相连的数据收集周期中，都要计算标识变量，并组成一个显示当前样本与累积样本数值的报告表。逐月计算的质量表被当成是监督数据收集过程中的一致性的基础。使用这些数据，如有必要，可以实施纠正收集过程的方案。

调查的一个目标是描述芬兰 6 岁以上的登记人口的移动情况。使用了按比例配额的分层简单随机抽样来抽取样本。分层是基于年龄、性别、地域。数据收集在时间上分为 12 个月，每月包括从中央人口登记抽取的 1 500 人。数据收集发生在 1998 年 7 月到 1999 年 6 月。调查涵盖了一整年的每一天，这样就涵盖了在时间上移动的所有变化。数据是逐月输入的。因此，有 12 个数据文件。

为了确保田园工作的质量，访谈员上岗前接受了训练，并按月监督数据的收集。访谈员也按时接到对他们工作的反馈，以便知道自己收集的资料是否有着连贯一致的质量。预期的应答者事先得到了关于调查的信息。比如，他们收到了一封描述调查背景与目标的信函。

我们首先给出 4 个关键质量标识的经验结果：涵盖率（%）、应答率（%）、访谈员效应以及离异系数（%）。然后，简要讨论两个逐月收集的资料质量的监督报告表中的一个。

监督涵盖率

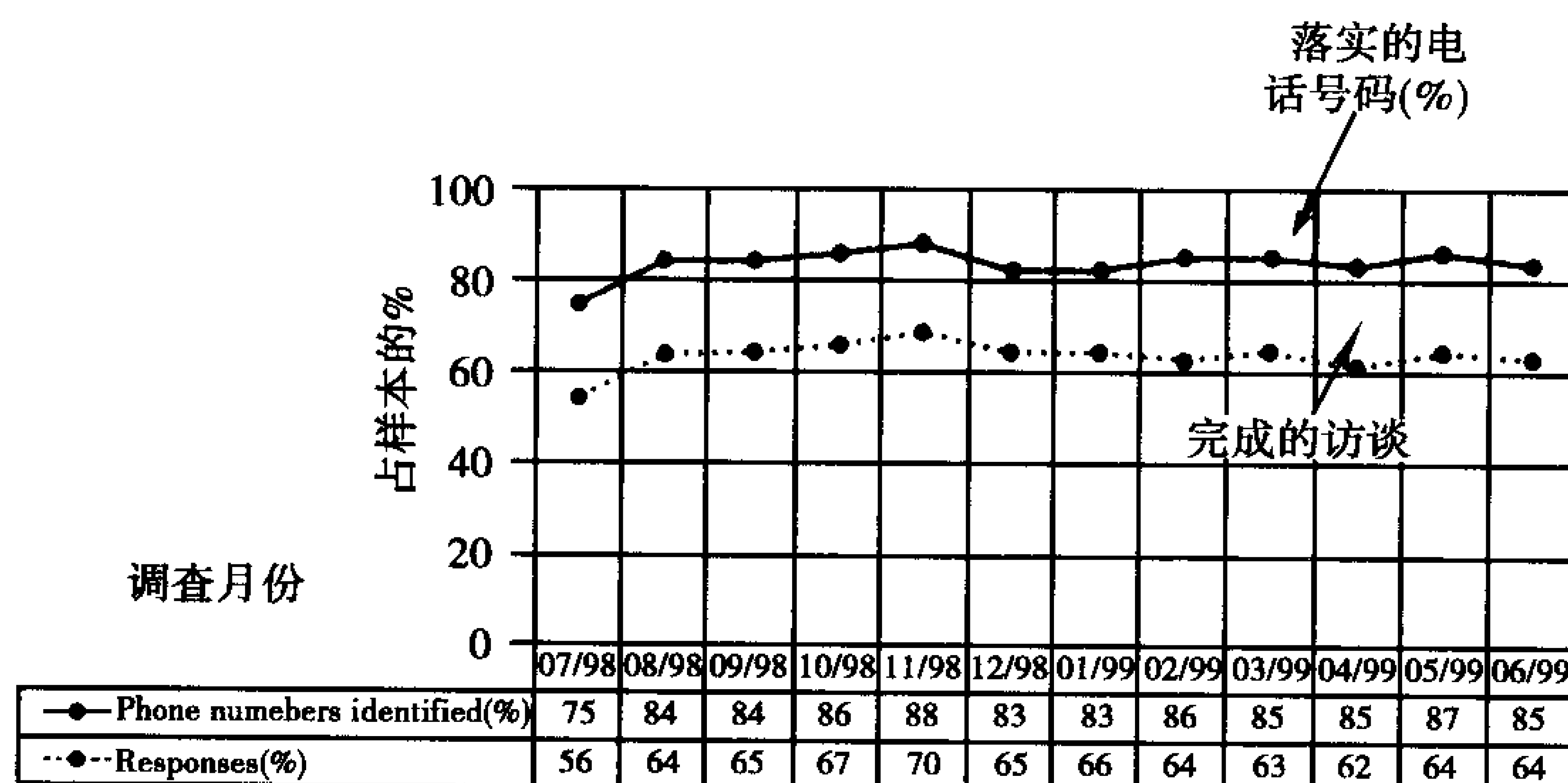
在这个调查中，涵盖率（%）的定义如下。抽样的总体框架由相关的人口登记构成。电话号码框架由电话号码登记与所有人姓名组成。当这两个登记并不一定相同时，涵盖误差就出现了。涵盖率的估计值为，

$$涵盖率(\%) = (n_F/n) \times 100$$

其中，n_F是电话号码在框架中标明的样本个人的数目，n 是样本规模。

电话普及率可以作为一个例子。在计算机辅助电话访谈中，目标人群可能是国内居住在家庭的所有成年人。框架总体——电话号码的数据库表格，仅包括可以由电话联系到的个人。通常，这个框架总体明显小于目标总体，因此形成涵盖不足的误差。这是一个由无法观察引起的非抽样误差。

根据库瑟拉（Kuusela，2000）报告，芬兰的普通电话覆盖率非常高。在1996年，超过96%的家庭拥有普通或是移动电话。高电话普及率，并不能确保电话访谈就能够有一个涵盖率高的数据收集。通常在寻找电话号码的过程中，会遇到相当多的问题。如图9.1所示，在乘客交通调查中，找到的电话号码比例为85%。因此，涵盖率不足的部分有15%。



	07/98	08/98	09/98	10/98	11/98	12/98	01/99	02/99	03/99	04/99	05/99	06/99
—◆— Phone numebers identified(%)	75	84	84	86	88	83	83	86	85	85	87	85
··◆·· Responses(%)	56	64	65	67	70	65	66	64	63	62	64	64

图9.1　各个调查月份中落实的电话号码与完成访谈的样本百分比

在电话访谈中，接触开始于落实现时的住址与电话号码信息。住址可以从新近的全国普查登记中挑选出来，但要找出电话号码则经常遇到问题。然而，即使一个家庭拥有电话号码，也不能保证能够找到它。

在1998年7月的第一个调查月里，找到的电话号码低于平均值。发现这一问题并作出调整之后，寻找电话号码的速度加快，在接下来的月份中结果得到了改善。

监督应答率

应答率（%）表示样本个人参与调查的比例。应答率的测量为，

$$应答率(\%) = \frac{I}{I + R + NC + O} \times 100,$$

式中，I——被访谈的人数；

R——合格的但拒访的人数；

NC——合格的但无接触的人数；

O——其他合格的但没有访谈的人数。

无应答的严重性有两方面：首先，它减少有效的样本规模，增大估计值的

标准误;第二,如果应答者对研究变量的赋值系统地偏离非应答者的数值时,造成可能的无应答偏差。因此,调查组织连续地记录无应答的原因(参见表 9.1)。

表 9.1 质量报告的例子:1999 年 6 月(1998—1999 年旅客交通调查)

指标	6 月	累计	备注
样本规模	1 500	18 250	按月/按年
落实的电话号码	84.7%	84.2%	涵盖率(%)
接触符合要求的人	77.3%	77.5%	接触率Ⅰ
接触在家的人	78.2%	77.9%	接触率Ⅱ
移动电话应答	16.1%	12.5%	接触率Ⅲ
完成访谈	63.9%	64.2%	应答率%
无法回答	0.9%	1.3%	无应答原因
语言问题	0.0%	0.2%	无应答原因
拒答及原因	12.5%	11.9%	无应答原因
没时间/忙	1.8%	2.0%	
不合作	5.5%	3.6%	
害怕个人资料泄漏	0.0%	0.0%	
无用的调查	0.2%	0.1%	
不知调查结果的用途	0.0%	0.0%	
无趣的调查	1.3%	1.6%	
其他原因	3.7%	4.6%	
访谈中断	0.1%	0.1%	无应答原因
无接触	22.7%	22.5%	无接触比率
旅行总数中已知目的地	76.3%	71.7%	测量误差
访谈员数目	10	19	按月/按年
每个访谈员完成的访谈	96	616	访谈员/工作量
旅行数的组内相关系数	0.071	0.017	访谈员效应
每天公里数的组内相关系数	0.002 4	0.001 6	访谈员效应
旅行数的离异系数	2.8%	0.8%	抽样误差
每天公里数的离异系数	9.9%	2.7%	抽样误差

来源:帕斯蒂能(Pastinen,1999)。1998—1999 年旅客交通调查(芬兰语)。交通与传播部出版物 43/99。芬兰:Edita Ltd。

从图 9.1 中可看出,除了 1998 年 7 月——调查的第一个月——以外,逐月计算的应答率为 65% 左右。在整个调查时期内,应答率时间上的差异并不显著。在乘客交通调查中,寻找电话号码与最后的应答率高度相关。如果只

找到了少量的电话号码，访谈员在提高应答率上能做的很少。这解释了1998年7月偏少的应答率(%)。由于1998年7月较低的电话号码寻找比例，其间的应答率比平均应答率小10%。与其他全国性的调查相比，这个调查的应答率并没有显著的不同。从无应答的角度，格罗夫斯等(Groves et al.,2001)报告了几个全国性的调查。

监督访谈员效应

访谈员效应(interviewer effect)属于非抽样误差。电话或是当面访谈，是一个访谈员与应答者之间的社会互动过程。比默等(Biemer,1991)列出了可能产生访谈员效应的4个方面：(a)调查访谈被看成是结构性的社会互动；(b)访谈员在填写问卷时有所不同；(c)不同的强调词与语气；(d)应答困难时的独特反应。所有这4个因素都可能产生访谈员的内部相关。评估这一调查误差的常用统计量是组内相关系数(Kish,1962)。用 $\overline{m}$ 来表示访谈员的平均工作量，组内相关可以由以下公式来估计

$$\hat{\rho}_{int} = \frac{\left(\frac{\hat{V}_b - \hat{V}_w}{\overline{m}}\right)}{\left(\frac{\hat{V}_b - \hat{V}_w}{\overline{m}}\right) + \hat{V}_w},$$

其中的访谈员方差组成部分 $\hat{V}_b$ 是以访谈员为因子的一元方差分析的组间平方差，$\hat{V}_w$ 是相应的组内平方差，其取值在 $-\frac{1}{m} \leqslant \hat{\rho}_{int} \leqslant 1$ 之间变动。注意，这个公式与前面第2章与第3章给出的系统抽样与整群抽样的公式不同。那里的组内相关是在基于设计的分析选择下定义的。这里的起点是由访谈员引起的测量误差的模型。因此，是在允许工作量变化的基于模型的分析选择下计算组内相关。访谈员效应引起的组内相关的影响，应当包括到估计值的标准误估计中。比如，如果遇到非0的 $\hat{\rho}_{int}$，估算的样本均值的设计方差应当乘上一个膨胀因子 $\text{deff} = 1 + (\overline{m} - 1)\hat{\rho}_{int}$。

接下来，给出乘客交通调查的经验发现。选择每人每天的移动次数作为一个研究变量。在图9.2中，给出了作为月份数字与累积数字的 $\hat{\rho}_{int}$。注意，每月的数值按各月的工作量分别计算；对于累积数值，则各个访谈员的各月的工作量合并在一起。前两个月(1998年6月与7月)没有数值，因为这一监督开始于1998年8月。

在累积数值上，$\hat{\rho}_{int}$ 平均值为0.02。许多研究结果表明，在大规模电话调查中的 $\hat{\rho}_{int} \approx 0.02$ 相当普遍(Groves,1989)。如果这个例子中有访谈员效应，它将增大置信区间，抵消增大样本规模而降低抽样误差的效应。这一结果的建议是，在大规模长时期调查中，应当限制每个访谈员的最大工作量，以避免

同一个访谈员访谈过多的应答者。因此,抽中的个人应当随机分给访谈员,这一做法可以在抽中个人中间化解访谈员效应(Biemer et al. ,1991)。

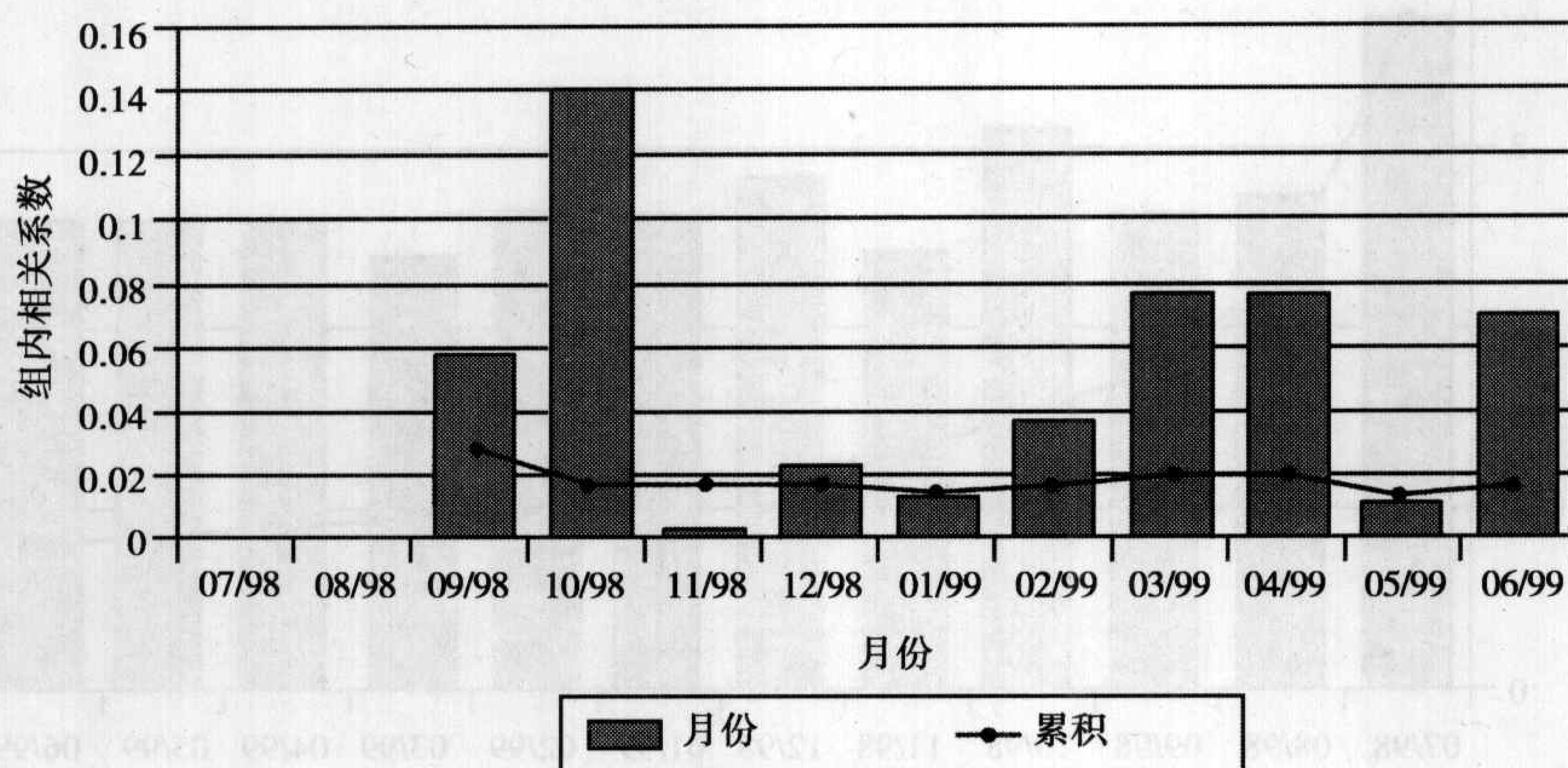


图 9.2　每人每天旅行次数的组内相关系数

以月份数值为基础,可以计算组内相关对于样本均值的平均膨胀效应。在 1999 年 6 月,$\hat{\rho}_{int}$为 0.071,访谈员的平均工作量 $\overline{m}$为 96 个应答者。所以,膨胀因子为,$deff = 1 + \hat{\rho}_{int}(\overline{m} - 1) = 1 + 0.071 \times (96 - 1) = 6.75$ 。比如,为了控制访谈员效应,估算的样本均值的标准误应当乘以这个因子的平方根,

$$s.e(\overline{y}) = \sqrt{6.75} \times s.e(\overline{y})_{p(s)} = 2.60 \times s.e(\overline{y})_{p(s)}。$$

使用离异系数监督抽样误差

由 $C.V(\hat{\theta})\%$ 表示的离异系数(coefficient of variation)测量相对抽样误差。对于一个非负的研究变量,点估计 $\hat{\theta}$ 的估算离异系数为 $c.v(\hat{\theta}) = s.e(\hat{\theta})/\hat{\theta}$。为了变量、调查以及逐月数据间的比较更加简便,离异系数被定义为以下百分比,

$$\text{COEFFICIENT OF VARIATION}(\%) = \frac{s.e(\hat{\theta})}{\hat{\theta}} \times 100。$$

图 9.3 给出了每人每天平均移动次数的离异系数的逐月与累积数值。累积数值清楚地表明,随着观测数目的增加,离异系数降低。

平均月份数值为 2.7%,显示了仅有较小的相对抽样误差。与期望的一致,随着月份数据的增加,c. v(%)的累积数值平稳下降。

质量报表的格式

调查组织单位决定了给出两份月度质量报告,以便监控连续的数据收集,并获得周期中的质量一致性。第一份表格,包括从月份数据与累积数据中计

算出来的25个不同的指标。表9.1给出了这种格式的例子。

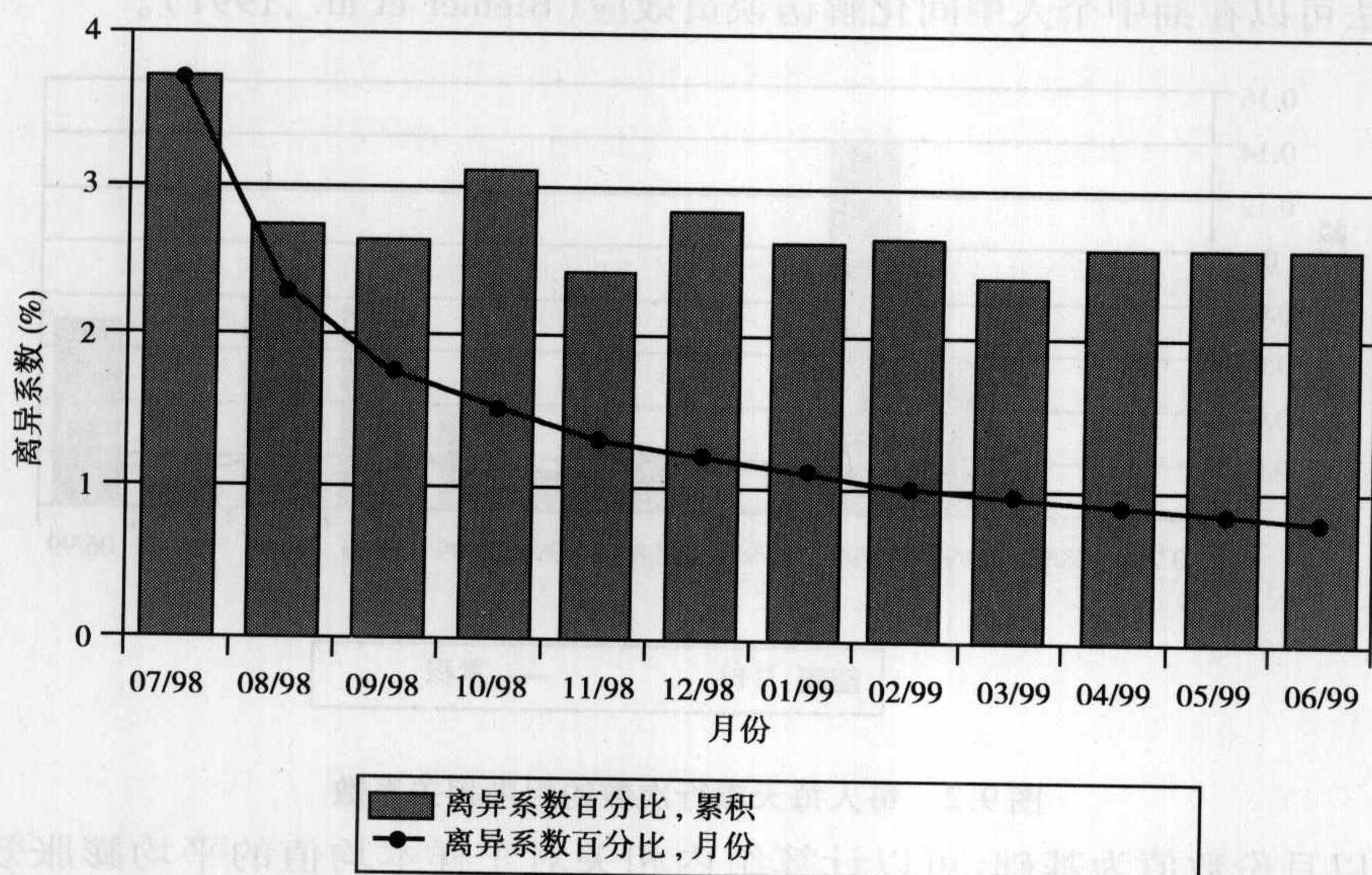


图9.3 每人每天的平均旅行次数的离异系数(%)，月份与累积数

这一报告的对象是客户与调查组织。在表9.1中，累积数值是月份数值的比较基础。比如，最后两行给出了离异系数。在实际中，估算所有感兴趣的变量的离异系数很重要。这对于检查是否超过了公布中期结果的最大可接受范围尤为重要。

9.2 商业调查中平均工资的估算

本例的关注点是，使用从商业公司收集来的资料，估算商业中不同职业雇员的平均工资。在抽样设计中，主要抽样单位是单个的公司。这意味着，在雇员层次的工资数据是以公司为整群。因此，在估算过程中要相应地考虑这样的设计。实际的抽样设计是一级分层整群抽样。在估算作为整体的商业部门的平均工资，以及本部门某些职业群体的平均工资时，为了便于比较，也使用了另外3个抽样设计分析选择。

抽样设计

抽样框架是一个商业登记表。其中，商业公司被分成两个子总体。第一个由商业雇主联合会(简写为CCE公司)的所有成员公司组成。从这个子总体中，联合会收集不同职业工资的普查资料。在下面的比较中，在完整数据基础上计算出来的平均工资将作为参照。

另一个子总体由不属于商业雇主联合会的公司组成。从这个子总体中，以个体公司为主要抽样单位，抽取一个分层简单随机样本。我们的目标是，使用收集的样本资料来估算这一子总体不同职业的平均工资。

在当前样本的抽样框架中，首先从商业登记表中剔除了最小的公司（只雇用 1 ~2 个人的）。这样，得到 25 345 个公司。根据雇员人数与商业分支，分别分成 5 个组别，共得到 25 个层级。不同的层级，抽样比例不同；在某些层级，所有公司均被抽中；而其他层级中，仅有一些公司进入样本。在商业登记表上，各个层级中单个公司的排列顺序是随机的。接下来，从上往下，从各个层级中抽取所需数目的公司。最初的样本规模为 1 572。去除 165 个已经涵盖的 CCE 公司，76 个不合格的公司，38 个已倒闭的公司，得到 1 369 个商业公司。回答的公司有 1 100 个，所以应答率为 80%。

到目前为止的抽样仅限于公司层次，抽样设计可以被描述为一个无放回式分层简单随机抽样。如果是在公司层次做结论，则可以在分层简单随机抽样设计中做分析。比如，这样的设计非常适合分析工作变化及与此相似的数据。

但是，这里的目的是估计不同职业的雇员的平均工资。这意味着抽样设计的不同解释。因为，作为分析单位的单个雇员，并不是主要抽样单位。选中特定的公司，意味着其中的所有雇员均被选中。因此，每个选中的公司应当被看成为一个整群，其中的元素是公司所有雇员。这样的样本设计被称为一级分层整群抽样（stratified one-stage cluster sampling）。抽样过程中仅有一个阶段，即是公司的抽取。在每个抽中的公司内，收集所有雇员的工资数据。

这里具体的对象是，商业领域在调查时间 1991 年 8 月的正式月工资。这些职业的分类是根据芬兰统计局的分类法。他们按时公布 22 个职业的平均工资。但有些类别太小，为了保密，仅列出了工作职位。这里的焦点是至少同时出现在 50 个抽样单位或是公司里的职业。一个明显特殊的数值是整个商业部门的平均工资。在当前样本中，由 744 个公司或是整群，13 987 个雇员组成。用抽样比例的倒数加权后，相应的规模为 $\hat{N}_{STATFIN}$ = 57 762 雇员。为了比较，CCE 登记表的雇员总数为 190 217。

加权与均值的估算公式

对于目前这种样本数据，可以根据抽样设计的假设，构建不同类型的均值估算公式。以下给出了 4 种不同的抽样设计，以及相应的均值与设计效应估计值。第 2,3 与 5 章给出了相应的方差估算公式，这里就忽略它们了。

1. 简单随机抽样。忽略公司层次，雇员层次的样本被当成直接从员工总体中得到的简单随机样本。因此，相应的平均工资的估算公式为，

$$\bar{y} = \frac{\hat{N}}{n}\sum_{k=1}^{n} y_k/\hat{N}, \tag{9.1}$$

其中，y_k 是样本中第 k 个雇员的工资。联合样本规模为 13 987。对所有雇员均使用相同的权重 $\hat{N}/n$；这是抽样比例倒数的近似值。权重为 $\hat{N}/n = 57\ 762/13\ 987 = 4.13$。只有当抽样发生在雇员层次，并且没有涉及分层与整群时，这个系数才是合理的。在当前的例子中，这些条件均不成立。在决定设计效应估计值———个总结设计复杂性对于方差估算的影响——时，均值估计值的方差是有用的。在第 2 章中，均值的设计效应被定义为两个方差估计值的比率：

$$\text{deff}(\bar{y}^*) = \frac{\hat{v}_{p(s)}(\bar{y}^*)}{\hat{v}_{srs}(\bar{y})}, \tag{9.2}$$

其中，$\bar{y}^*$ 是在实际抽样设计 $p(s)$ 下方差估计值为 $\hat{v}_{p(s)}(\bar{y}^*)$ 的均值估算公式，而 $\hat{v}_{srs}(\bar{y})$ 是在 SRSWOR 下 $\bar{y}$ 的方差估计值。如果设计效应接近于 1，实际抽样设计可以被当成一个 SRS 设计。这个例子中的分析并不需要抽样设计标识。在使用了整群抽样的例子中，设计效应可能大于 1。那样的话，为得到合适的分析，需要特殊的使用合适设计标识的软件。在 SRS 设计中，由定义，设计效应等于 1。

2. 分层简单随机抽样。假定了元素层次的抽样，并且每个层级有自己的权重。平均工资的估算公式为

$$\bar{y}_{str} = \sum_{h=1}^{H}\sum_{k=1}^{n_h}\frac{\hat{N}_h}{n_h}y_{hk}/\hat{N}。 \tag{9.3}$$

分层级的权重是 $\hat{N}_h/n_h$，或者是层级 h 中抽样比例的倒数。其中，$\sum_{h=1}^{H}\hat{N}_h = \hat{N}$，$\sum_{h=1}^{H}n_h = n$。值得注意，同一层级中的所有雇员的权重相同，即使（实际中也是如此）他们在不同的公司工作。

3. 层级权重不同的分层整群抽样。均值估算公式与分层简单随机抽样中的相同。但是，这一设计的用来决定置信区间的标准误估算公式不同。在分层整群抽样中，取决于不同研究变量的整群内部的同质性，其设计效应通常大于 1（$\text{deff} \geq 1$）。

4. 整群权重不同的分层整群抽样。这是一个符合商业公司样本实际情况的假设。由雇员数目表示的公司规模（即，整群规模），通常相差不小。在这种情形下，通过使用霍维茨-汤普森估算公式，以及使用整群的相对规模作为抽样权重，将设计纳入均值估算中。这里，整群的相对规模是，公司雇员数目 N_{hi} 除以相应层级中的雇员总数 N_h。这样将生成一个公司的整群权重。其倒数就是这个公司的抽样权重。为了使得权重的和与框架总体中的雇员总数相符，这一数值应当除以层级中的样本公司数目 m_h。因而得到均值估算公式为，

$$\bar{y}_{clu} = \sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{k=1}^{n_{hi}} \frac{\hat{N}_h}{m_h \times N_{hi}} y_{hik} / \hat{N}。 \tag{9.4}$$

这个估算公式纳入了所有抽样设计的信息:分公司与层级变化的抽样权重。

结　果

在分析样本数据时,合适地考虑了样本设计的特征。4 种不同的抽样设计假设下的估算,其加权的方案不同。它们在不同程度上考虑同一种抽样设计。显然,其中最合乎实际的设计假设是各个整群权重不同的分层整群抽样,它结合了所有抽样设计的信息。而 SRS 则是最简单的分析选择。这些抽样设计的结果,也可以与从 CCE 普查中得到的平均工资统计量相比较。在表 9.2中,CCE 的数据显示在最后一行。芬兰统计局样本显示,估算的雇员人数为 57 762。这意味着,在 1991 年 8 月,本部门的全职雇员总数为 57 762 + 190 217 = 247 979。

表 9.2　1991 年商业部门雇员的平均工资(欧元)(根据不同抽样设计的假设及普查资料)

抽样设计	加权样本规模	平均工资	标准误	deff
SRS	57 762	1 759	7.4	1.00
STR(分层)	57 762	1 602	9.3	1.72
CLU(分层加权)	57 762	1 602	10.1	2.10
CLU(整群加权)	57 762	1 581	11.1	2.58
普查(CCE 登记表)	190 217	1 530	0	n. a.

n. a. :不适用。

SRS 设计的估计值给出了最大的平均工资 1 759 欧元。另一方面,它同时又有最小的标准误估计值 s. e = 7.4。在其他设计中,平均工资接近于普查中得到的参照值,1 530 欧元。由于这是相应子总体的准确数值,它没有标准误。与参照值最接近的估计值来自整群权重不同的分层整群抽样设计,其平均工资的估计值为 1 581 欧元。在这个设计中,主要抽样单位是公司,但是在雇员层次上加权。

结果的比较

接着去看商业职业群体的平均工资。表 9.3 比较 3 种来源的数值:商业雇主联合会的登记数据、基于分层一级简单随机抽样的芬兰统计局估计值,以及最后的从整群权重不同的分层整群抽样中得到的估计值。比较仅包括了至少有 50 个公司同时含有的最大的职业类别(表 9.3)。

表 9.3 1991 年 8 月不同职业的平均工资:CCE 会员公司的普查与芬兰统计局样本

职业群体	1991 年 8 月平均工资		
		芬兰统计局样本	
	CCE 普查	CLU 设计	STR 设计
商店经理	1 612	1 486	1 430
服务站工人	1 159	1 173	1 161
清洁工	1 150	911	906
仓库工人	1 195	1 196	1 191
面包车司机	1 313	1 201	1 216
运输代理商	1 504	2 164	2 293
其他部门	1 414	1 288	1 303
上层白领	2 545	2 427	2 421
办公室经理	3 231	3 306	3 326
办公室管理人员	2 349	2 523	2 542
办事员	1 494	1 708	1 707
司机	1 613	1 332	1 324
所有职业	1 530	1 581	1 602

基于普查数据的数值与基于芬兰统计局样本的数值之间有差异。但是，由于仅仅只有较少职业群体中有这样的差异，仔细地检查不同统计来源使用的职业分类的内部相容性，则更加有用。总体而言，从整群权重不同的分层整群抽样中得到的估计值，比基于分层简单随机抽样假设的芬兰统计局估计值更接近于普查数值。

完整设计信息的使用，显著提高了平均工资估计值的标准误。一个可能的原因是，在编辑抽样框架与具体抽样时点之间，公司的规模在原有的规模类别间有所变动，但没有改变该层级的权重。从芬兰统计局使用的样本设计的设计效应中，可以清楚地显示这一点。分公司的权重有两个方面的影响。第一，由于考虑了实际的规模，它减轻了前面所讲的框架过时的问题。第二，它们引入了一个群内正相关的整群效应，因此，使用整群权重不同的分层整群抽样，增加了平均工资的标准误及相应的设计效应。

结　论

这个例子演示了混合类型的数据收集策略。商业公司的目标总体有两个部分：雇主联合会的会员登记表，以及没有登记的公司。为了生成可靠的工资统计量，需要各个公司工资的信息，因而对整个商业行业有强烈的应答压力。数据的主要来源是普查类型的行政登记表。对于余下的公司或是未登记的公

司,芬兰统计局以公司为单位,用分层简单随机抽样抽取了样本。因此,整个商业行业中,仅有抽中的公司填写问卷。这一程序极小化了这种调查带来的额外应答压力。另一方面,正如我们所演示的,应当使用不同的估算策略,小心分析这种设计收集的数据。

整群设计的相对较高的设计效应估计值($2.10 \leqslant deff < 2.58$),进一步显示,整群效应不可忽视,在计算商业公司的平均工资时,应当考虑到这一点。整群效应是指,在同一公司相同职业岗位上的人(如,店铺职员),工资差不多相同。但他们的工资与其他公司相同职业的人却不同。这一结果也建议,计算平均工资时,应当使用整群层次的权重。另一个支持整群层次权重的因素是,公司(整群)规模的变动范围较大。最自然的方法是使用霍维茨-汤普森估算公式。考克斯等(Cox et al.,1995)总结了近年来在商业调查方法上的进展。

9.3 社会经济调查中的模型选择

在这个例子中,我们演示了不仅仅是考虑到整群效应很关键,模型的设定与预测变量的假设可能也很重要。为了这一目标,我们将章节 8.3 与章节8.4 介绍的通用加权最小二乘法(GWLS)与类似然方法(PML),应用于组群比例的比率对数 ANOVA 与 ANCOVA 模型中。在这个演示中,我们使用 3 种分析选择(参见章节 8.2)。基于设计的分析选择(分析选择 1)考虑了这个例子中所有的抽样设计复杂性,即加权与整群。加权 SRS 分析选择(分析选择 2)假定简单随机抽样,但考虑到了加权。未加权分析选择(分析选择 3)假定简单随机抽样,并忽略所有的抽样复杂性。研究的问题是评价一个疾病保险计划。数据是芬兰健康保障调查抽样设计中的一个地域层级。其中,家庭户为整群,并作了无应答的调整。

研究问题与数据

疾病保险的一个重要目的是,减少不同人口群体使用健康服务的差异,以及降低疾病给个人与家庭带来的财务负担。在芬兰,从 1964 年以来,就有了一个覆盖整个人口的公共疾病保险计划。在 1980 年代,一个由私有保险公司提供的补充疾病保险计划的使用越来越广。这一计划的例子如返还因到私有健康保障部门看医生治病的费用。我们将使用芬兰健康保障(FHS)调查中的数据,来研究各个收入群体在使用私有保险上的差异。这个调查由芬兰社会保险研究院在 1987 实施。

FHS 调查的目的是为评估健康与社会保障提供可靠的信息。使用了地域

性的一级分层整群抽样。实际情况与数据收集的经济因素,促成了使用家庭户作为数据收集单位。在6 998 个样本家庭中,5 858 个家庭(84%)参与了这次调查。样本家庭中的所有合格的成员构成了元素层次的样本,共访谈了16 269非机构收养的人。无应答集中在城市地区,特别是诸如赫尔辛基的大城市。由于不可忽视的无应答问题,使用了后续分层,后续层级由地域、性别与年龄组构成。

分家庭户进行了个人访谈,但主要的兴趣却是在个人层次的推论。显然,在家庭内部,与健康相关的特征——健康服务的使用、保健行为等,同质性非常高。所以,相应的研究变量可能是正的群内相关。这些变量的均值与比例估计值的设计效应通常将大于1,但小于2。设计效应最大的估计值(deff = 1.7)是测量是否使用私有疾病保险的二分变量 INSUR。

这个例子中使用调查中的子样本,它由来自赫尔辛基都市的878 个家庭与2 071 个人组成。它是35 个层级中的一个。赫尔辛基地区使用私有疾病保险的人数的比例估计值较高,为17%;这里的私有健康服务的供给也比其他地区高。在农村,这一比例相当低。

检查 INSUR 与家庭收入之间的联系,与评价公共疾病保险计划相关。但是,初步分析并不支持参与私有疾病保险取决于高收入的假设。在3 个家庭收入类别(低、中、高)中,估算的 INSUR 比例分别为15.2%,17.3%与18.1%。这些比例的同质性检验中,得到了皮尔逊检验统计量的观测值为 $X_P^2 = 2.15$, p-值为0.342,显然是不显著的差异。另外,以 INSUR 为回应变量,家庭收入为整数取值1,2,3 的定量预测变量的对数回归中,p-值为0.148,显示了线性趋势并不显著。

但是,是否参加私有健康保险极大地取决于年龄。看起来,私有保险是特别为小孩使用的疾病保险形式。在赫尔辛基都市地区,43%的小孩为这种保险覆盖,而成年人的比例仅有9%。另外,因为慢性病与急性病而有访问医生的需要,增加参加私有保险的概率。对于那些在一段时间内至少访问医生一次的人群中,参见私有疾病保险的比例为27%;而另一人群的比例是14%。当然,因果关系(如果有的话)也可能是另一个方向。因此,当更仔细地研究家庭成员参与私有保险与家庭收入的关系时,将应答者年龄与访问私人医生当成一个合成因素,可以提供更多信息。

为了进一步研究联系,交叉分类表格的 ANOVA 类型的比例对数模型给出了最简便的方法。为了简便,我们选择了二分变量 VISITS(在固定的时期内至少访问私人医生1 次)与 AGE(0~17 岁的孩子;或是17 岁以上的成年人),以及3 个组别的变量 INCOME(每个 OECD 消费者单位的家庭收入,1/3部分)作为 ANOVA 模型中的预测变量。在这些预测变量下,共生成12 个人口子群或是组群。由于 INCOME 可以被当成一个定量的预测变量,我们也用

这些变量拟合 ANCOVA 模型，以便进一步检查对 INCOME 是否有线性趋势。

表 9.4 给出了 INSUR 的组群比例。比例 $\hat{p}_j^U = n_{j1}/n_j$，INSUR 的组群样本和 n_{j1} 以及组群样本规模 n_j 都是原有的、忽略了加权的 SRS 分析选择下的未加权数量值。在其他两种分析选择下，使用了对无应答在加权的比例 $\hat{P}_j = \hat{n}_{j1}/\hat{n}_j$。因此，比例估计值是一致的比率估计值。其中的 $\hat{n}_{j1}$ 与 $\hat{n}_j$ 分别是加权的组群样本和与加权的组群样本规模。设计效应估计值 $\hat{d}_j$ 对应加权比例估计值 $\hat{p}_j$。同时，也给出了样本整群数目 m_j，即各个子群覆盖的家庭户数。

表 9.4 赫尔辛基都市区私人参保人数(INSUR)未加权与加权的比例估计值 $\hat{P}_j^U$ 及 $\hat{P}_j$(%)，分 VISITS，AGE 及 INCOME(FHS 调查)

组群	VISITS	AGE	INCOME	$\hat{p}_j^U$	n_j	$\hat{p}_j$	$\hat{d}_j$	$\hat{n}_j$	m_j
1	无	小孩	低	27.6	145	29.0	1.7	140	86
2			中	33.3	135	33.6	1.7	125	93
3			高	41.3	75	41.2	1.3	69	57
4		成人	低	6.7	400	6.5	1.5	422	258
5			中	8.9	427	8.6	1.5	425	245
6			高	11.6	423	11.3	1.6	422	256
7	有时	小孩	低	60.5	43	60.3	1.4	44	33
8			中	74.4	39	75.2	1.4	37	30
9			高	75.6	41	75.4	1.3	41	35
10		成人	低	12.6	103	12.9	1.3	110	92
11			中	12.5	88	11.4	1.0	87	83
12			高	11.2	152	10.5	1.3	149	127
整个样本			17.2	2 071	16.8	1.8	2 071	878	

INSUR：参加私营疾病保险(二项回应)。

VISITS：在给定时期内至少造访私人医生一次。

AGE：年龄(小孩 0～17 岁/成人 18 岁及以上)。

INCOME：每经合组织消费者单位的 1986/1987 年家庭净收入(1/3 部分)。

固定 VISITS 与 AGE，除了最后 3 个收入组以外，INSUR 的比例随着收入的增加而增加。总体上讲，VISITS 中的第 2 组与 AGE 的第 1 组中的比例要大一些。至少访问医生 1 次的孩子组的比例最大。设计效应显示了较小的整群效应，其均值为 1.4。

方 法

首先，以 VISITS，AGE 与 INCOME 为定性预测变量，对 INSUR 比例 $\hat{p}_j$ 与

$\hat{p}_j^U$ 用 GWLS 方法拟合对数 ANOVA 模型。接着,用 PML 方法对相同的表格拟合 ANCOVA 模型,并将 INCOME 当成取值为 1 到 3 的定量变量。我们在章节 8.2 中介绍了 3 种分析选择下使用 GWLS 与 PML 方法。在未加权 SRS 分析选择下,忽略所有设计复杂性;在加权 SRS 分析选择下,只考虑权重的问题;在基于设计的分析选择下,允许了额外的二项协变与不同比例估计值间的相关。最后一种分析选择使用了实际的整群抽样设计,另外两种分析选择假定了简单随机抽样。

有明显的原因,支持基于设计的分析。如果一个家庭成员——特别是孩子,参加了保险,其他人则也会参加。那么,回应变量 INSUR 则含有正的群内相关。这一整群效应,可以由整体上的 INSUR 比例的设计效应估计值 deff = 1.8 显示出来。清楚显示了额外的二项协变的组群设计效应,也证实了这一点。

就组群结构而言,还有一个关于群内相关的重要问题。显然,VISITS 与 AGE 横切整群,并组成交叉分类,即家庭。INCOME 则组成相互区隔的类别,因为它是一个家庭层次的预测变量。这些预测变量一起构成了一个混合类别的结构。这导致了相互区隔的比例 $\hat{p}_j$ 间的成对相关。并不是所有比例都是相关的,只有那些相对于各自 INCOME 组——即,每一个第 3 组——的才有相关。所以,除了二项协变以外,这些组群中的比例估计值也应当有正的协方差。这也支持使用基于设计的分析方法。

群内相关的结构反映在组群比例 $\hat{p}_j$ 的 12×12 的协方差估计值 $\hat{\mathbf{V}}_{des}$ 中。图 9.4给出了这一估计值。其中,为了比较,也给出了二项估计值 $\hat{\mathbf{V}}_{bin}$。由于自由度较大,$f = m - H = 877$,用线性化方法得到的估计值 $\hat{\mathbf{V}}_{des}$ 相当稳定,其条件数也并不大(37.4)。因此,可以认为在基于设计的分析选择下,GWLS 与 PML 方法较为满意。因为 $\hat{\mathbf{V}}_{des}$ 对角线上的方差估计值大于相应的二项方差估计值,相对于在基于设计分析选择下得到的检验结果,在 SRS 分析选择下的结果更大。

如第 8 章所示,根据分析选择,GWLS 与 PML 方法的对数模型需要比例估计值向量及其协方差矩阵估计值。在 GWLS 分析中,使用了章节 8.3 中等式 8.5 到式 8.13;在 PML 分析中,使用了章节 8.4 中等式 8.24 到式 8.27。在基于设计的分析选择下,使用了估计值 $\hat{p}_j$ 与 $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}})$;在加权 SRS 分析选择下,除了 $\hat{p}_j$ 以外,使用了 $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}})$;在未加权 SRS 分析选择下,使用了未加权估计值 $\hat{p}_j^U$ 与 $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}}^U)$ 。

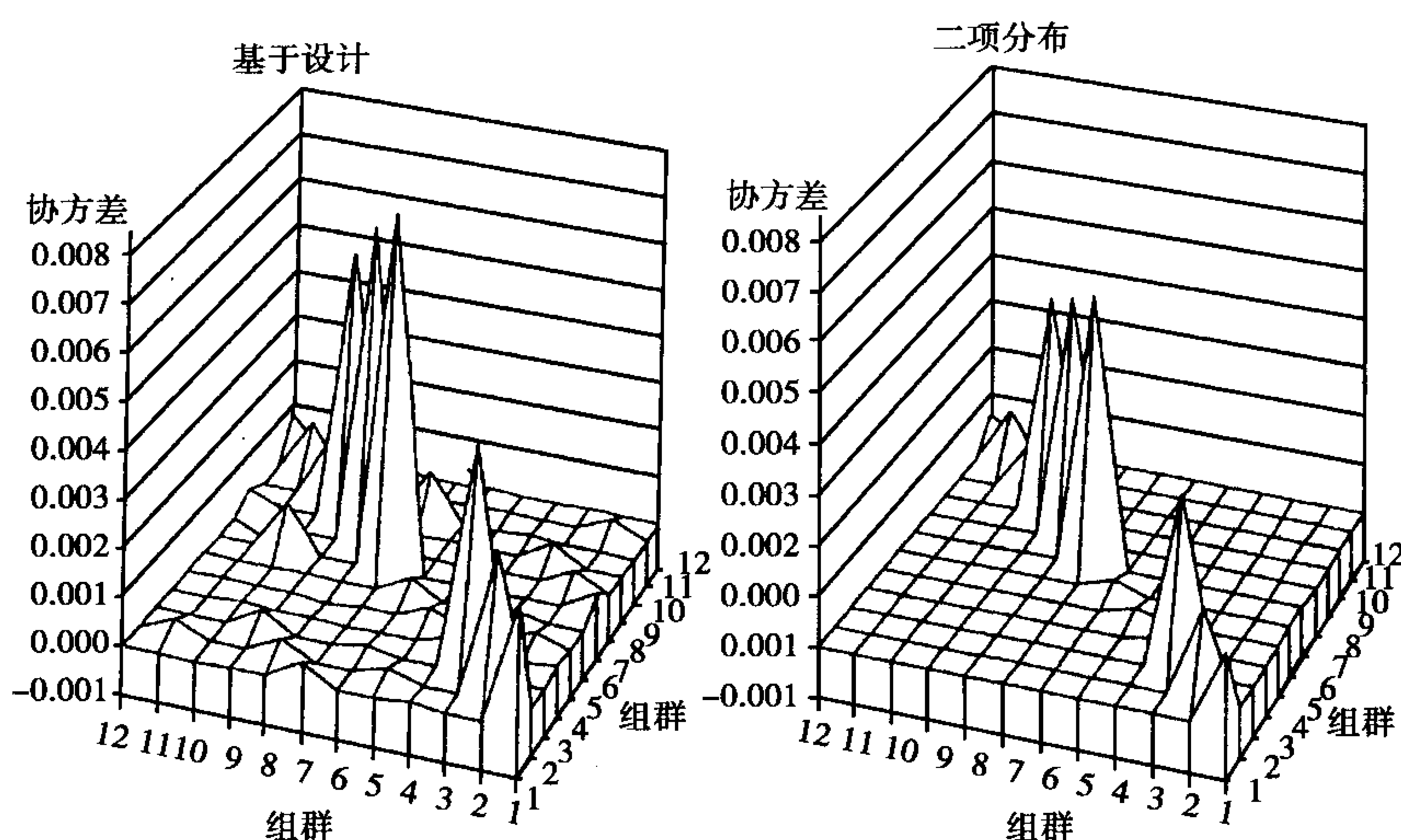


图 9.4 INSUR 比例 $\hat{p}_j$ 的协方差矩阵估算。基于设计的估算 $\hat{\mathbf{V}}_{des}$ 与二项分布估算 $\hat{\mathbf{V}}_{bin}$（FHS 调查）

结　果

让我们首先讨论比率对数 ANOVA 模型的检验结果。我们希望研究，在控制访问医生与应答者年龄的符合效应的基础上，参加私有保险与家庭收入之间的独立性。除了相应的主效应以外，需要检查可能的交互效应。所以，完全模型的形式为 $\log[(P/1-P)] = V + A + I + V*A + V*I + A*I + V*A*I$。其中，V 表示 VISITS，A 表示 AGE，I 表示 INCOME，以及 P 表示参加私有保险的组群比例。注意，在这种表达式中，所有预测变量都被当成是定类的。

含有所有主效应以及 VISITS 与 AGE 的交互效应的模型得到较为满意的拟合，并且无法进一步简化。表 9.5 的靠左部分给出模型拟合度的结果，包括基于 SRS 的沃尔德统计量的观测值，以及基于设计的沃尔德统计量。由于样本数目较大，没有必要调整不稳定的 F-校正。在任何分析选择下，根据检验结果，简化的模型拟合较好。

分析的主要兴趣在于，作为参加私有保险的预测变量，ANOVA 模型中 INCOME 效应的重要性。表格的中间部分给出了成分分析假设下的沃尔德检验结果，用 $X^2(\mathbf{b})$ 表示。检验结果显示，在基于 SRS 分析选择下，INCOME 的效应非常显著。最宽松的检验——在 1% 水平显著，发生在未加权的 SRS 分析选择下。在加权 SRS 分析选择下，在 5% 水平显著。这两个检验均忽略了整群效应。但是，一旦使用基于设计的分析选择，考虑二项协变与组群比例的

相关时,INCOME 的效应变得并不显著。INCOME 的效应即使在 10% 水平也不显著。

表 9.5 基于设计与基于 SRS 分析选择下,对数 ANOVA 模型拟合度,INCOME 效应显著性,及低对高的 INCOME 比照,显著性的沃尔德检验结果(FHS 调查)

分析选择	模型拟合			INCOME 效应显著性			低对高比照的显著性
	X^2	df	p-值	X^2(**b**)	df	p-值	p-值
选择 1	4.23	6	0.645 0	4.35	2	0.113 8	0.037 2
选择 2	4.52	6	0.606 3	7.95	2	0.018 8	0.004 8
选择 3	3.61	6	0.729 0	9.31	2	0.009 5	0.002 3

选择 1:在实际整群抽样设计下,基于设计的分析。

选择 2:简单随机抽样假设,加权分析。

选择 3:简单随机抽样假设,未加权分析。

为了更详细的推论,我们分别检验低收入组与高收入组的模型参数相等的假设。对相应的"低对高"比照的检验结果出现在表 9.5 靠右的部分。所有的检验显示差异至少在 5% 水平显著,而且 p-值的变化模式与前面一致。

接下来,我们使用估算的模型系数及其标准误(表 9.6)计算相应的修正的比率比及其 95% 的置信区间。我们计算了极端的分析选择 1 与 3。

在这两个分析选择下,INCOME 的第 1 组的修正比率比显著(5% 水平)不等于 1——高收入组的比率比。

虽然在考虑整群效应时,整体上家庭收入看起来并不显著,但比率对数 ANOVA 模型的结果有些支持在两个极端收入组之间参与私有疾病保险的可能性并不相等的结论。因此,理应进一步建模来清楚地检验在控制混合因素后 INCOME 组别间可能的线性趋势。使用 ANCOVA 模型来达到这一目标。其中,INCOME 被当成一个定量预测变量,整数值 1 到 3 被赋予到其各个组别。这样,我们提高了对变量 INCOME 固有信息的使用。

现在,使用 PML 方法来拟合对数 ANCOVA 模型。含有前面 ANOVA 模型中相同模型项的模型的结果显得比较合理,但需要进一步的检查。

表 9.6 基于设计与未加权 SRS 分析选择下，INSUR 的修正比率比统计量(FHS 调查)

分析选择	比率比	比率比 95% 的置信区间	
		下限	上限
选择 1			
INCOME 等级			
1	1	1	1
2	1.22	0.81	1.85
3	1.56	1.03	2.38
选择 3			
INCOME 等级			
1	1	1	1
2	1.23	0.91	1.69
3	1.64	1.19	2.22

选择 1:在实际整群抽样设计下,基于设计的分析。

选择 3:简单随机抽样假设,未加权分析。

让我们更详细地讨论这个模型中 INCOME 的回归系数 b_4 的检验结果。表 9.7给出了在基于设计与未加权 SRS 分析选择下得到的结果。事实上,未加权 SRS 的结果是基于 ML 方法,因为忽略了加权。在这个表中,t-检验结果显示,INCOME 的回归系数显著与 0 不同(5% 水平)。同时,基于 SRS 分析选择的检验比基于设计的检验更大。加权 SRS 分析选择下的检验结果介于它们之间。还要注意,估计值 $\hat{b}_4$ 有些不同,但加权 SRS 分析选择下将得到与基于设计分析选择下相等的估计值。

表 9.7 基于设计与未加权 SRS 分析选择下,用 PML 方法拟合对数 ANOVA 模型的 INCOME 回归系数(b_4)的估算与检验结果(FHS 调查)

分析选择	$\hat{b}_4$	$\hat{d}(\hat{b}_4)$	s.e$(\hat{b}_4)$	t 检验	p-值
选择 1	0.229	1.77	0.109	2.10	0.035 7
选择 3	0.246	1.00	0.081	3.02	0.002 6

选择 1:在实际整群抽样设计下,基于设计的分析。

选择 3:简单随机抽样假设,未加权分析。

小 结

我们研究了在控制访问医生与应答者的年龄的复合效应后,参加私有疾病保险是否取决于家庭收入。为了分析,数据被安排成组群比例的多维表格。

比例显示了轻微的整群效应。对数 ANOVA 模型给出了研究比例差异的最简单方法。当忽略了整群效应时,家庭收入的效应看起来是显著的。

一旦考虑到这些效应时,它就失去了显著性。在检验一个对比时,以及比率比的估计值中,有些结果又显示了,在参加私有疾病保险上,两个极端收入组有差异。因而,需要进一步解释。在对数 ANCOVA 模型中,更加明确地检验了家庭收入效应的线性趋势。结果显示,高收入是决定是否参加私有疾病保险的重要因素。同时显示,作为公共保险计划的补充,私有保险计划涉及了参加与使用健康保健服务的不平等。

在前面的分析中,参与私有疾病保险的变量是二分回应变量。这主要是为了演示的目的。这一变量的群内相关相对较强。将健康服务用作回应变量,保险变量作为其中一个预测变量,也将是合理的。这样的话,可能开启问题的另一个视角。

方法上的结论

正如这个例子所示,即使相关相对较弱,回应变量的正的群内相关,可能严重地扭曲多变量分析中的检验结果。在对数 ANOVA 与 ANCOVA 模型中,忽略整群效应比恰当地考虑整群效应得到的检验结果相对要大些。这是因为,忽略整群效应时,低估了系数标准误。因此,这样的结果,对于在整群抽样得到的数据中使用标准分析方法而言,是一种警示。对于与这里的分析相关的干扰方法,使用了因素加权的、最小二乘法或是似然法的基于设计的估算,为分析群内相关的回应变量提供了一个安全而又容易控制的方法。同样的,这样的结果也应当与其他模型形式相比较,以便得到关于这类问题有效的推论。

9.4 教育调查中的多级建模

在研究学生扫盲的多国教育调查中,对垂直分级结构数据中的连续回应变量使用了多层级模型。作为整群的学校,反映出了总体的分级结构,并以它进行了整群抽样。抽样设计对回应变量产生了较强的群内相关。在分析中,需要将这一特征考虑在其中。这里介绍的分解方法,给出了一个不同于本书主要讲解的干扰或是聚合方法的形式。我们使用分解方法,对若干国家的数据分别拟合一个两层级的线性模型。结果也将与忽略了设计复杂性的分析相比较。

PISA:一个国际教育调查

数据来自于 OECD 的国际学生评估计划(PISA)。第一次 PISA 调查于

2000 年在 28 个 OECD 国家以及 4 个非 OECD 国家实施。2000 年的 PISA 调查涵盖了 3 个主题范围:读写扫盲、数学扫盲以及科学扫盲。这里,我们讨论读写扫盲。从 PISA 调查中,我们选择了以下国家:巴西、芬兰、德国、匈牙利、韩国、英国与美国。国家选择是有讲究的。这些国家既有地域上的代表性,其整群效应也各不相同。从 7 个国家得到的调查数据共有 1 388 个学校,32 101 个学生。

2000 年的 PISA 调查使用了一个高度结构化的调查设计,包括基本概念、程序、测量工具——如问卷、抽样设计、数据收集过程,以及估算与分析过程的标准化。这样是为了尽可能地确保结果间的国际比较可能性。

学校与学生的抽样

在教育调查的抽样设计中,很自然地使用现存的学校行政性与功能性结构。可以将学校管理区域或是其他行政标准分组的学校当成基本单位。另一方面,教学也分成有学生与老师组成的教学单位或是班级。在教育调查中,因为经济与其他实际因素,学校通常被当成数据收集的主要单位。从抽中的学校中,选取学生作为二级单位。因此,总体中有自然的分级结构。在这个例子的抽样设计与模型分析阶段,都将使用这一特征。

在 PISA 国家中,大多使用了二级分层整群抽样。第一级由注册了 15 岁学生的单个学校组成。抽取学校时,使用了系统 PPS 抽样(参见章节 3.2),规模是估算的注册合格学生(15 岁)人数。在绝大多数国家中,学校在抽样前被分层。在每个国家,至少抽取了 150 个学校(给定有这么多的学校情况下)。在国别分析时,通常要求比这个数字更大的样本。

在第二阶段,在抽中的学校内抽取学生。在抽取的学校内,准备了一个 15 岁学生的名单。从这个名单上,以等概率抽取了 35 个学生。如果少于 35 个注册学生,则抽取所有学生。

抽取的学校至少要达到 85% 的应答率;抽中的学生至少要达到 80% 的参与率。这一最低参与率是在国家层次,而非单个的学校(OECD,2001,2002a)。

加权方案

对于每个国家的样本数据,建立了合适的抽样权重。元素权重反映了以下因子:学校的选中概率、学校内学生的选中概率以及学校与学生的无应答修正。对于每个国家,学校 i 中学生 k 的权重 w_{ik} 可以表示如下:

$$w_{ik} = w_{1i} \times w_{2ik} \times f_i, \quad i = 1, \cdots, m \text{ 及 } k = 1, \cdots, n_i,$$

其中,$w_{1i} = 1/(\pi_i \hat{\theta}_i)$ 是学校 i 的选中概率 π_i 与估算的参与概率 $\hat{\theta}_i$ 的乘积的倒数;

$w_{2ik} = 1/(\pi_{k|i} \hat{\theta}_{k|i})$ 是学校 i 中学生 k 的选中概率 $\pi_{k|i}$ 与估算的参与概率 $\hat{\theta}_{k|i}$ 的乘积的倒数;

f_i 是补偿调查设计中可能的分国家的限制在学校 i 中的修正因子;

m 是给定国家中的样本数目，n_i 是学校 i 中的样本学生数目。

经过换算，用于分析中的学生层次的因素权重，加总起来等于各个国家样本数据的实际规模。在给定的国家中，换算后的权重的均值为 1，但在不同的国家，权重有所不同。换算后权重的最小标准差为 0.143，最大的为 0.983。OECD(2002b)给出了更为详细的加权程序。

部分国家的识字率

结果变量 y 是学生的合成识字率取值（或者，更准确的，5 个可信的合成识字率取值中的第 1 个）。经过换算，得到 OECD 国家的共同均值为 500，标准差为 100。我们称回应变量为合成识字率取值。表 9.8 给出了部分国家识字率的描述性统计量。我们使用了第 5 章中的技术，计算了合成识字率的均值与标准误。因此，这些估计值是基于设计的，并恰当地考虑了各个国家中抽样设计的复杂性（加权、分层与整群）。表中有两种不同的设计效应。总体设计效应包括加权、分层与整群。第二个设计效应包括分层与整群，而允许在加权 SRS 分析选择下的比较。两个设计效应显示，在大多数国家中整群效应较强。在某些国家中，两个设计效应的差别相当可观，显示权重的差异较大。

表 9.8　PISA2000 年调查合成识字率的描述性统计量（分国家；字母排序）

国别	合成识字率					数据中的个案数	
	均值	标准误	整体设计效应	由分层与整群产生的设计效应	学生的有效样本规模	学生	学校
巴西	402.9	3.82	8.33	5.17	476	3 961	290
芬兰	550.7	2.15	2.79	2.74	1 600	4 465	147
德国	497.4	5.68	13.47	11.68	305	4 108	183
匈牙利	485.7	6.02	20.00	16.20	231	4 613	184
韩国	526.6	3.66	12.99	11.67	351	4 564	144
英国	531.4	4.08	14.08	7.16	564	7 935	328
美国	517.0	5.16	6.93	5.46	354	2 455	112

数据来源：经合组织 PISA 数据库，2001。

将学生人数除以总体上的设计效应，就计算出了学生的有效样本规模。学生的有效样本规模等于使用简单随机抽样而不使用整群时满足相同的估算精度所需的样本规模。当观测个案并不相互独立，有效样本规模就下降了：设计效应越高，有效样本规模就越小。虽然，所有国家中名义上的样本规模都较大（几千），但其中某些国家的有效样本却较小（仅有几百）。

设计效应估计值也显示，在错误的简单随机抽样的假设下计算的标准误，将大大小于大多数国家的基于设计的标准误估计值。

拟合两级线性模型

在分析中，结果变量 y 是合成识字率取值。解释结果变量的变动由 2 个

学校层次与 4 个学生层级的变量来完成。学校层次的解释变量是学校规模(SSIZE)与老师自主性(AUTONOMY)。学校规模即是学校里的实际学生人数除以 100。校长被问到,在学校的几种事务中,谁负主要责任。老师自主性由校长认为老师负主要责任的事务数目中转换得来。两个变量均被标准化,使得 OECD 国家的共同均值为 0,标准差为 1。

学生层次的解释变量是,学生的性别(编码是 1 为女生,0 为男生,变量名为 FEMALE)、社会经济背景(SEB)、阅读练习(ENGAGEMENT),以及成绩压力(ACHPRESS)。SEB 指标从学生回应父母职业中得出;阅读练习指标,从学生回答关于阅读习惯与态度的问题中得出;而成绩压力指标,从学生报告的来自老师的压力得出。同样的,这 3 个指标也经过了标准化,所有 OECD 国家的共同均值为 0,标准差为 1。

含有两个层级的解释变量与随机变动的、合成识字率取值的两级回归模型如下:

$$\begin{aligned} y_{ik} = {} & \text{INTERCEPT} + \gamma_1 \times \text{SSIZE}_i + \gamma_2 \times \text{AUYONOMY}_i + \\ & \beta_1 \times \text{FEMALE}_{ik} + \beta_2 \times \text{SEB}_{ik} + \beta_3 \times \text{ENGAGEMENT}_{ik} + \\ & \beta_4 \times \text{ACHPRESS}_{ik} + u_i + e_{ik}, \end{aligned}$$

其中,指数 k 对应层级 1 单位(学生),i 对应层级 2 单位(学校)。固定效应 γ 与 β 分别表示学校与学生层次变量的回归系数。残差 u_i 是学校 i 的随机效应,假定是均值为 0、方差为 σ_u^2 的正态分布。而 e_{ik} 是学生层级的残差,假定是均值为 0、方差为 σ_e^2 的正态分布。随机效应 u_i 与 e_{ik} 假定相互独立。分析中使用了学生层级的换算权重。

就所关注的变量而言,与从总体中随机选取相比,现存自然的整群——如学校——中的因素间更为相似或是同质。这意味着,层级 1 中的因素(学生)在统计上不能被假定在学校内相互独立,因而研究变量是正的群内相关。在多层级模型中,用如下公式来估算群内相关(skinner et al. ,1989;Coldstein,2002;Snijders and Bosker,2002):

$$\hat{\rho}_{int} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}^2},$$

其中,估算的研究变量的总体方差 $\hat{\sigma}^2$ 被分成两个部分,学校间的方差 $\hat{\sigma}_u^2$ 与学校内的方差 $\hat{\sigma}_e^2$。群内相关系数测量相同层级 2 组别(学校)内层级 1 单位(学生)数值间的成对相关,并被称为校内相关系数。在基于模型的分析选择中,从空心模型——两个层级仅含截距与残差的多层级模型——的方差部分来估算这个系数。比如,表 9.9 中,匈牙利的校内相关系数的估算为,6 093.7/(6 093.7 +3 148.3) =0.659。也可以从含有解释变量模型的方差成分来估算这一系数。这里,它被称为残差校内相关系数。表 9.10 中,匈牙利的残差校内相关系数的估算为,4 744.2/(4 744.2 +2 897.4) =0.621。注意,本书前面(章节 3.2)在基于设计的分析选择中,使用了群内相关的概念。

表 9.9 PISA2000 年调查合成识字率两级方差成分模型（空心模型）估算值（分国家；按校内残差相关系数大小排序）

国别	校内残差相关系数	方差成分		截距	标准误
		学校层级	学生层级		
匈牙利	0.659	6 093.7	3 148.3	464.1	5.84
德国	0.553	5 572.2	4 507.8	496.1	5.61
巴西	0.428	3 146.9	4 201.4	387.9	3.61
韩国	0.375	1 828.6	3 043.0	520.9	3.74
美国	0.241	2 318.2	7 35.5	503.3	4.97
英国	0.212	1 917.5	7 126.5	529.0	2.88
芬兰	0.063	470.7	6 960.9	550.6	2.18

资料来源：经合组织 PISA 数据库，2001。

表 9.10 PISA2000 年调查合成识字率两级模型估算值（分国家）

		匈牙利	德国	巴西	韩国	美国	英国	芬兰
固定效应：								
系数								
截距	γ_0	471.2	496.4	382.0	506.8	496.6	524.9	531.6
	s.e	6.36	4.58	4.56	6.29	6.05	3.38	4.91
	t-检验	74.14	108.37	83.75	80.53	82.12	155.06	108.27
	p-值	0.000	0.000	0.000	0.000	0.000	0.000	0.000
学校层级变量：								
学校规模	γ_1	30.6	27.4	2.4	7.1	1.0	3.8	5.9
	s.e	9.00	9.22	1.47	3.44	2.54	3.14	7.35
	t-检验	3.41	2.94	1.64	2.07	0.38	1.20	0.80
	p-值	0.001	0.003	0.100	0.039	0.705	0.232	0.426
教师自主性	γ_2	4.8	−7.1	−3.1	2.5	4.1	−2.3	2.8
	s.e	5.62	5.22	4.24	5.39	3.63	2.61	2.68
	t-检验	0.86	−1.37	−0.74	0.47	1.14	−0.89	1.06
	p-值	0.392	0.171	0.459	0.641	0.256	0.374	0.291
学生层级变量：								
女	β_1	6.4	3.6	3.1	15.9	14.9	9.8	19.6
	s.e	2.22	2.41	2.54	2.49	3.71	2.64	2.43
	t-检验	2.89	1.50	1.21	6.38	4.00	3.71	8.09
	p-值	0.004	0.133	0.228	0.000	0.000	0.000	0.000
社会经济背景	β_2	6.0	11.5	9.9	2.2	16.7	23.3	15.8
	s.e	1.09	1.53	1.35	0.92	2.22	1.32	1.34
	t-检验	5.56	7.50	7.34	2.40	7.51	17.70	11.78
	p-值	0.000	0.000	0.000	0.016	0.000	0.000	0.000

续表

		匈牙利	德国	巴西	韩国	美国	英国	芬兰
阅读练习	β_3	19.5	19.0	19.5	16.6	28.9	31.5	33.9
	s.e	1.04	0.98	1.51	1.04	1.99	1.40	1.26
	t-检验	18.68	19.36	12.87	15.94	14.49	22.59	27.05
	p-值	0.000	0.000	0.000	0.000	0.000	0.000	0.000
成绩压力	β_4	0.9	-1.6	3.4	3.4	-3.3	-7.2	-3.7
	s.e	0.93	1.16	1.44	0.89	2.04	1.59	1.40
	t-检验	0.92	-1.35	2.36	3.85	-1.62	-4.52	-2.65
	p-值	0.356	0.176	0.018	0.000	0.106	0.000	0.008
随机效应：								
方差成分								
学校层级		4 744.2	3 501.6	2 730.5	1 387.3	1 770.6	999.6	394.8
学生层级		2 897.4	3 981.9	3 830.6	2 809.6	6 094.1	5 779.0	4 984.3
校内残差相关系数		0.621	0.468	0.416	0.331	0.225	0.147	0.073
与空心模型相比，方差减少的比例(%)								
学校层级		22.1	37.2	13.2	24.1	23.6	47.9	16.1
学生层级		8.0	11.7	8.8	7.7	16.7	18.9	28.4
合计		17.3	25.8	10.7	13.8	18.4	25.0	27.6

资料来源：经合组织 PISA 数据库，2001。

方差的成分通过受限最大似然方法(REML)来估算。给定这些方差估计值，固定效应则通过通用最小二乘法(GLS)来估算(Bryk and Raudenbush，1992)。与此同时，也得到了消除整群效应的标准误(参见，比如章节 8.4 的“夹逼”形式)。

表 9.9 给出了基本的两级方差成分模型，即，没有解释变量的空心模型。这些模型中，估算了一个固定效应——截距，以及学校层级的随机截距。总体上的方差被分成校间与校内两个部分，并将它们用来计算校内相关系数。这些国家中，估算的系数差异较大，最小值为 0.063，最大值为 0.659。

给定一个国家，表 9.9 中的截距是估算的平均学校截距。这些截距与表 9.8 中的各个国家均值有些不同。估算的截距的标准误估计值也有所不同。因为，在计算它们的过程中，使用了估算出的多级模型。

表 9.10 给出了估算的合成识字率取值的两级模型。在学校层级，学校规模的效应在某些国家统计上显著。第 2 个学校层级的变量——老师自主性，其效应在所有国家中统计上均不显著。

在学生层级的解释变量中，社会经济背景与阅读练习的效应，对于每个国家，至少在 5% 水平统计上显著。混合经济背景的效应在国家间差异很大。社会经济背景越好，他或她更多地练习阅读，则他或她的阅读熟练数值就越高。成绩压力效应的强度与方向差异很大。在大多数国家中，性别效应在统

计上显著。

由其减少的比例所示，相当数量学校与学生层级的识字率差异为估算的模型所解释。但是，与空心模型相比，拟合模型在降低程度上也增加了可观的差异。在大多数国家中，与未解释的总体差异相比，学校层级的未解释的差异仍然较大。这可以从残差校内相关系数的数值上看出。

模型中仅包括了解释变量的线性效应。也可以研究某些变量（如，学校规模）可能的抛物线性的效应。所有层级 1 解释变量的系数被当成是固定效应。但是，系数中可能也有学校间的差异。如果是那样的话，也可以使用随机系数回归模型。

与加权 SRS 分析相比较

我们最后将多层级模型得到的结果与忽略整群效应而得到的结果相比较。我们使用对应于观测个案互相独立假设的加权 SRS 分析（参见章节 8.2）。在这种分析选择下，使用与两级模型相似的解释变量，对结果变量拟合一个固定效应的线性模型。我们选择德国的数据做比较（表 9.11）

德国数据的回应变量群内相关很高。因此，在加权 SRS 分析选择下拟合模型中，估算的固定层级 2 效应的标准误估计值过小。如果使用加权 SRS 分析分析选择，层级 2 效应之一——老师自主性，将错误地被认为在统计上显著。而学校规模的效应将被估算得太小。层级 1 的解释变量中，与两级模型的结果相比，社会经济背景与阅读练习的效应过大。而成绩压力也成为了一个统计上显著的效应。

表 9.11　合成识字率的两级模型与加权 SRS 分析选择的固定效应模型的系数估算比较（以德国数据为例）

系　数		两级模型	加权 SRS 分析选择
截　距	γ_0	496.4	497.5
	s.e	4.58	1.93
	t-检验	108.37	258.08
	p-值	0.000	0.000
学校规模	γ_1	27.4	20.1
	s.e	9.22	1.74
	t-检验	2.97	11.52
	p-值	0.003	0.000
教师自主性	γ_2	−7.1	−7.3
	s.e	5.22	1.38
	t-检验	1.37	5.26
	p-值	0.171	0.000
女	β_1	3.6	3.3
	s.e	2.41	2.74
	t-检验	1.50	1.20
	p-值	0.133	0.229
社会经济背景	β_2	11.5	31.5
	s.e	1.53	1.38
	t-检验	7.50	22.9
	p-值	0.000	0.000

续表

系 数		两级模型	加权 SRS 分析选择
阅读练习	β_3	19.0	28.9
	s.e	0.98	1.17
	t-检验	19.36	24.6
	p-值	0.000	0.000
成绩压力	β_4	-1.6	-4.7
	s.e	1.16	1.31
	t-检验	-1.35	-3.64
	p-值	0.176	0.000

资料来源:经合组织 PISA 数据库,2001。

小 结

这个例子显示,由于忽略了观测个案间的正的群内相关,对于以整群抽样得到的数据,假定观测个案间的独立性而做的分析,可能是严重误导的。仅仅在没有整群效应的假设下,两级模型的结果与加权 SRS 分析的结果相似。

我们使用了“分解”方法,用两级模型明确地对多层级结构的总体建模。分析垂直分层结构数据的另外的方法是第 8 章讨论的基于设计的方法。在那里,与对垂直分层结构建模不同,数据结构带来的整群效应被当成是干扰。因此,在基于设计的分析中,我们试着在估算与检验结果中“剔除”整群效应,以获得有效推论。

从实际的角度来看,多层级模型的额外贡献在于,它明确地给出了整群间差异的信息,因此,为解释结果给出了更多的信息。



Bean J. A. (1975) DiStribution and properties Of variance estimators for complex multistage probabilitv samples *Vital and Health Statistics* Series 2, NO. 65.

Biemer P. P., Groves R. M., Lyberg L. E., Mathiowetz N. A. and Sudman S. (eds) (1991) *Measurement Errors in Surveys Chichester*: Wiley.

Biemer P. P. and Lyberg L. E. (2003) *Introduction to Survey Quality* New York: Wiley.

Binder D. A. (1983) On the Variances Of asymptotically normal estimators from complex surveys *International Statistical Review* **51** 279-292.

Binder D. A. (1991) A framework for analyzing categorical survey data With non-response *Journal Of Official Statistics* **7** 393-404.

Binder D. A. (1992) Fitting Cox's proportional hazards models from survey data *Biometrika* **79** 139-147.

Breslow N. E. and Clayton D. g. (1993) Approximate inference in generalized linear mixed models *Journal Of the American Statistical Association* **88** 9-25.

Brewer K. R. W. (1963) Amodel Of systematic sampling With unequal probabilities *Australian Journal of Statistics* **5** 5-13.

Brewer K. R. W. and HanifM. (1983) *Sampling with Unequal Probabilities* New York: Springer.

Brier S. S. (1980) Analysis of contingency tables under cluster sampling *Biometrika* **67** 591-596.

Bryk A. S. and Raudenbush S. W. (1992) *Hierarchical Linear Models: Applications and Data Analysis Methods* Newbury Park: Sage Publications.

Chambers R. and Skinner C. (eds) (2003) *Analysis of Survey Data* Chichester: Wiley.

Clayton D., Spiegelhalter D., Dunn G. and Pickles A. (1998) Analysis Of longitudinal binary data from multiphase sampling *Journal of the Royal Statistical Society, B* **60** 71-87.

Cochran W. G. (1977) *Sampling Techniques* Third Edtion. New York: Wiley.

Couper M., Baker R., Bethlehem J., Clark C., Martin J., Nicholls II W. and O'Reilly J. (eds) (1998) *Computer Assisted Survey Information Collection* New York: Wiley.

Cox B. G., Binder D. A., Chinnappa B. N., Christiansson A., Colledge M. J. and Kottp. S. (eds) (1995) *Business Survey Methods* New York: Wiley.

Datta G. S., Lahiri P., Maiti T. and Lu K. L. (1999) Hierar chical Bayes estimation of unemployment rates for the states of the U. S. *Journal of the American Statistical Association* **94** 1074-1082.

Dempster A. P., Rubin D. B. and Tsutakawa R. K. (1981) Estimation in covariance component models *Journal of the American Statistical Association* **76** 341-353.

Deville J.-C. and Särndal C. E. (1992) Calibration estimators in survey sampling *Journal of the American statistical Association* **87** 376-382.

Deville J.-C., Särndal C. E. and Sautory O. (1993) Genralized raking procedures in survey

sampling *Journal of the American Statistical Association* **88** 1013-1020.

Diggle P. J. , Heagerty P. J. , Liang K. -Y. and Zeger S. L. (2002) *Analysis of Longitudinal* Data Second Edition Oxford : Oxford University Press.

Dillman D. (1999) *Mail and Internet Surveys : The Tailored Design Method* Second Edition New York : Wiley.

Efron B. (1982) *The Jackknife , The Bootstrap and Other Resampling Plans* Philadelphia : Society for Industrial and Applied Matematics.

Estevao V. , Hidiroglou M. A. and Särndal C. -E. (1995) Methodological principles for a generalized estimation system at Statistics Canada *Journal of Official Statistic* **11** 181-204.

Estevao V. M. and Särndal C. -E. (1999) *The use of auxiliary information in design-based estimation for domains* Survey Methodology **25** 213-221.

Feder M. , Nathan G. and Pfeffermann D. (2000) Multilevel modelling of complex survey longitudinal data with time varying random effects *Survey Methodology* **26** 53-65.

Federal Committee on Statistical Policy (2001) Measuring and Reporting Sources of Error in Surveys *Statistical Policy Working Paper* 31 , Washington DC : Statistical Policy Office , Office of Management and Budget.

Fellegi I. P. (1980) Approximate tests of independence and goodness of fit based on stratified multistage samples *Journal of the American Statistical Association* **75** 261-268.

Francisco C. A. and Fuller W. A. (1991) Quantile estimation with a complex survey design. *Annals of Statistics* **19** 454-469.

Frankel M. R. (1971) *Inference from Survey Samples* Ann Arbor : Institute for Social Research , The University of Michigan.

Freeman D. H. (1988) Sample survey analysis : analysis of variance and contingency tables. In : Krishnaiah P. R. and Rao C. R. (eds) *Handbook of Statistics* 6. *Sampling.* Amsterdam : North Holland , 415-426.

Ghosh M. (2001) Model-dependent small area estimation : theory and practice. In : Lehto-nen R. and Djerf K. (eds) *Lecture Notes on Estimation for Population Domains and Small Areas* Helsinki : Statistics Finland Reviews 2001/551-108.

Ghosh M. and Natarajan K. (1999) Small area estimation : a Bayesian perspective. In : Ghosh S. (ed.) *Multivariate Analysis , Design of Experiments , and Survey Samplig* New Yok : Marcel Dekker , 69-92.

Ghosh M. , Natarajan K. , Stroud T. W. F. and Carlin B. (1998) Generalized linear models for small area estimation *Journal of the American Statistical Association* **93** 273-282.

Ghosh M. and Rao J. N. K. (1994) Small area estimation : an appraisal *Statistical Science* **9** 55-93.

Glynn R. J. , Laird N. M. and Rubin D. B. (1993) Multiple imputation in mixture models for nonignorable nonresponse with follow-ups *Journal of the American Statistical Association* **88** 984-993.

Goldstein H. (1987) *Multilevel Models in Educational and Social Research* London : Griffin.

Goldstein H. (1991) Nonlinear multilevel models , with an application to discrete response data *Biometrika* **78** 45-51.

Goldstein H. (2002) *Multilevel Statistical Models* Third Edition London : Edward Arnold.

Goldstein H. and Rasbash J. (1992) Efficient computational procedures for the estimation of parameters in multilevel models based on iterative generalized least squares *Computational Statistics and Data Analysis* **13** 63-71.

Grizzle J. E. , Starmer C. F. and Koch G. G. (1969) Analysis of categorical data by linear models *Biometrics* **25** 489-504.

Groves R. M. (1989) *Survey Errors and Survey Costs* New York : Wiley.

Groves R. M., Dillman D. A., Eltinge J. L. and Little R. J. A. (2001) *Survey Nonrespones* New Yok: Wiley.

Hansen M. H. and Hurwitz W. N. (1943) On the theory of sampling from a finite population *Annals of Mathematical Statistics* **14** 333-362.

Hedayat A. S. and Sinha B. K. (1991) *Finite Population Sampling* New Yok: Wiley.

Heliövaara M., Aromaa A., Klaukka T., Knekt P., Joukamaa M. and Impivaara O. (1993) Reliability and validity of interview data on chronic diseases *Journal of Clinical Epidemiology* **46** 181-191.

Hidiroglou M. A. and Rao J. N. K. (1987a) Chi-squared tests with categorical data from complex surveys: Part I *Journal of Official Statistics* **3** 117-132.

Hidiroglou M. A. and Rao J. N. K. (1987b) Chi-squared tests with categorical data from complex surveys: Part II *Journal of Official Statistics* **3** 133-140.

Holt D., Scott A. J. and Ewings P. D. (1980) Chi-squared tests with survey data *Journal of the Royal Statistical Society, A* **143** 303-320.

Holt D. and Smith T. M. F (1979) Post stratification *Journal of the Royal Statistical Society, A* **142** 33-46.

Holt D., Smith T. M. F. and Tomberlin T. J. (1979) A model-based approach to estimation for small subgroups of population *Journal of the American Statistical Association* **74** 405-410.

Horton N. J. and Lipsitz S. R. (1999) Review of software to fit generalized estimating equation regression models *The American Statistician* **53** 160-169.

Horvitz D. G. and Thompson D. J. (1952) A generalization of sampling without replacement from a finite universe *Journal of the American Statistical Association* **47** 663-685.

Judkins D. (1990) Fay's method for variance estimation *Journal of Official Statistics* **6** 223-240.

Kalton G. (1983) *Introduction to Survey Sampling* Beverly Hills: Sage Publications.

Keyfitz N. (1957) Estimates of sampling variance where two units are selected from each stratum *Journal of the American Statistical Association* **52** 503-510.

Kish L. (1962) Studies of interviewer variance for attitudinal variables *Journal of the American Statistical Association* **57** 92-115.

Kish L. (1965) *Survey Sampling* New York: Wiley.

Kish L. (1992) Weighting for unequal P_i *Journal of Official Statistics* **8** 183-200.

Kish L. (1995) Methods for design effects *Journal of Official Statistics* **11** 55-77.

Kish L. and Frankel M. R. (1970) Balanced repeated replications for standard errors *Journal of the American Statistical Association* **65** 1071-1094.

Kish L. and Frankel M. R. (1974) Inference from complex samples (with discussion) *Journal of the Royal Statistical Society, B* **36** 1-37.

Koch G. G., Freeman D. H. and Freeman J. L. (1975) Strategies in the multivariate analysis of data from complex surveys *International Statistical Review* **43** 59-78.

Korn E. L. and Graubard B. I. (1999) *Analysis of Health Surveys* New York: Wiley.

Krewski D. and Rao J. N. K. (1981) Inference from stratified samples: properties of the linearization, Jackknife and balanced repeated replication methods *Annals of Statistics* **9** 1010-1019.

Kumar S. and Singh A. C. (1987) On efficient estimation of unemployment rates from labour force survey data *Survey Methodology* **13** 75-83.

Kuusela V. (2000) *Telephone coverage situation in Finland.* (In Finnish). Helsinki: Statistics Finland, Reviews 3/2000.

Lawson A. B., Biggeri A., Böhning D., Lesaffre E., Viel J.-F. and Bertollini R. (eds) (1999)

Disease Mapping and Risk Assessment for Public Health Chichester: Wiley.

Levy P. S. and Lemeshow S. (1991) *Sampling of Populations: Methods and Applications* New York: Wiley.

Liang K. -Y. and Zeger S. L. (1986) Longitudinal data analysis using generalized linear models *Biometrika* **73** 13-22.

Liang K. -Y. Zeger S. L. and Qaqish B. (1992) Multivariate regression analyses for categorical data (with discussion) *Journal of the Royal Statistical Society, B* **54** 3-40.

Little R. J. A. (1991) Inference with survey weights *Journal of Official Statistics* **7** 405-424.

Little R. J. A. (1993) Post-stratification: a modeler's perspective *Journal of the American Statistical Association* **88** 1001-1012.

Little R. J. A. and Rubin D. B. (1987) *Statistical Analysis with Missing Data* New York: Wiley.

Lehtonen R. (1988) The Execution of the National Occupational Health Care Survey Helsinki: Publications of the Social Insurance Institution, Finland, M: 64. (In Finnish with English summary.)

Lehtonen R. (1990) On Modified Wald Statistics (Doctoral Dissertation). Their application to a Goodness of Fit Test of Logit Models under Complex Sampling Involving Ill-Conditioning Helsinki: Publications of the Social Insurance Institution, Finland, M: 74.

Lehtonen R. and Kuusela V. (1986) Statistical efficiency of the Mini-Finland health sur-vey's sampling design. Part 5. In: Aromaa A., Heliövaara M., Impivaara O., Knekt P. and Maatela J. (eds) *The Execution of the Mini-Finland Health Survey* Helsinki, Turku: Publications of the Social Insurance Institution, Finland, ML: 65. (In Finnish with English summary.)

Lehtonen R., Särndal C. -E. and Veijanen A. (2003) The effect of model choice in estimation for domains, including small domains *Survey Methodology* **29** 33-44.

Lehtonen R. and Veijanen A. (1998) Logistic generalized regression estimators *Survey Methodology* **24** 51-55.

Lehtonen R. and Veijanen A. (1999) Domain estimation with logistic generalized regression and related estimators. *Proceedings, IASS Satellite Conference on Small Area Estimation*, Riga, August 1999; Riga: Latvian Council of Science, 121-128.

Lohr S. L. (1999). *Sampling: Design and Analysis* New Yok: Duxbury Press.

Lundström S. and Särndal C. -E. (2002) *Estimation in the presence of Nonresponse and Frame Imperfections Statistics Sweden.* Örebro: SCB-Tryck.

Marker D. (1999) Organization of small area estimators using a generalized linear regression framework *Journal of Official Statistics* **15** 1-24.

McCarthy P. J. (1966) Replication. An approach to the analysis of data from complex surveys *Vital and Health Statistics* Series 2, No. 14.

McCarthy P. J (1969) Pseudoreplication: further evaluation and application of the balanced half-sample technique *Vital and Health Statistics* Series 2, No. 31.

McCarthy P. J. and Snowden C. B. (1985) The booststrap and finite population sampling *Vital and Health Statistics* Series 2, No. 95.

McCullagh P. and Nelder J. A. (1989) *Generalized Linear Models* Second Edition London: Chapman & Hall.

McCulloch C. E. and Searle S. R. (2001) *Generalized, Linear, and Mixed* Models New Yokrk: Wiley.

Morel J. G. (1989) Logistic regression under complex survey designs *Survey Methodology* **15** 203-223.

Moura F. A. S. and Holt D. (1999) Small area estimation using multilevel models *Survey*

Methodology **25** 73-80.

Murthy M. N. (1957) Ordered and unordered estimators in sampling without replacement *Sankhya* **18** 379-390.

Nathan G. (1988) Inference based on data from complex sample designs. In: Krishnaiah P. R. and Rao C. R. (eds) *Handbook of Statistics* 6. *Sampling* Amsterdam: North Holland, 247-266.

Nelder J. A. and Wedderburn R. W. M. (1972) Generalized linear models *Journal of the Royal Statistical Society, A* **135** 370-384.

OECD(2001) *Knowledge and Skills for Life* First results from the OECD Programme for International Student Assessment (PISA) 2000. Paris: OECD.

OECD (2002a) PISA 2000 Technical Report Paris: OECD(http://www.pisa.oecd.org/).

OECD (2002b) Manual for the PISA 2000 Database Paris: OECD.

Ohlsson E. (1998) Sequential Poisson sampling *Journal of Official Statistics* **14** 149-162.

Pastinen V. (1999) *Passenger Transport Survey* 1998—1999 (*In Finnish*) *Helsinki*: Publications of the Ministry of Transport and Communications, 43/99.

Pfeffermann D. (1993) The role of sampling weights when modeling survey data *International Statistical Review* **61** 317-337.

Pfeffermann D., Skinner C. J., Goldstein H., Holmes D. J. and Rasbash J. (1998) Weighting for unequal selection probabilities in multilevel models (With discussion) *Journal of the Royal Statistical Society. B* **60** 23-40.

Plackett R. L. and Burman J. P. (1946) The design of optimum mulitfactorial experiments *Biometrika* **33** 305-325.

Platek R. and Särndal C. -E. (2001) Can a Statistician Deliver? (With discussion) *Journal of Official Statistics* **17**, 1-127.

Prasad N. G. N. and Rao J. N. K. (1999) On robust small area estimation using a simple random effects model *Survey Methodology* **25** 67-72.

Quenouille M. H. (1956) Notes on bias in estimation *Biometrika* **43** 353-360.

Rao J. N. K. (1997) Developments in sample survey theory: an appraisal *The Canadian Journal of Statistics* **25** 1-21.

Rao J. N. K. (1999) Some recent advances in model-based small area estimation *Survey Methodology* **25** 175-186.

Rao J. N. K. (2003) *Small Area Estimation* New York: Wiley.

Rao J. N. K., Hartley H. O. and Cochran W. G. (1962) A simple procedure of unequal probability sampling without replacement *Journal of the Royal Statistical Society, B* **24** 482-491.

Rao J. N. K. and Scott A. J. (1981) The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables *Journal of the American Statistical Association* **76** 221-230.

Rao J. N. K. and Scott A. J. (1984) On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data *Annals of Statistics* **12** 46-60.

Rao J. N. K. and Wu C. F. J. (1985) Inference from stratified samples: second-order analysis of three methods for nonlinear statistics *Journal of the American Statistical Association* **80** 620-630.

Rao J. N. K. and Scott A. J. (1987) On simple adjustments to chi-square tests with sample survey data *Annals of Statistics* **15** 385-397.

Rao J. N. K. and Thomas D. R. (1988) The analysis of cross-classified categorical data from complex sample surveys *Sociological Methodology* **18** 213-269.

Rao J. N. K. and Wu C. F. J. (1988) Resampling inference wiht complex survey data *Journal of the American Statistical Association* **83** 209-241.

Rao J. N. K. and Thomas D. R. (1989) Chi-squared tests for contingency tables. In: Skin-

ner C. J. ,Holt D. and Smith T. M. F. (eds) *Analysis of Complex Surveys* Chichester: Wiley, 89-114.

Rao J. N. K. Kumar S. and Roberts G. (1989) Analysis of sample survey data involving categorical response variables: methods and software(With discussion) *Survey Methodology* **15** 161-186.

Rao J. N. K. and Scott A. J. (1992) A simple method for the analysis of clustered binary data *Biometrics* **48** 577-585.

Rao J. N. K. , Wu C. F. J. and Yue K. (1992) Some recent work on resampling methods for complex surveys *Survey Methodology* **18** 209-217.

Rao J. N. K. and Shao J. (1993) Jackknife variance estimation with survey data under hot deck imputation *Biometrika* **79** 811-822.

Rao J. N. K. , Sutradhar B. C. and Yue K. (1993) Generalized least squares *F* test in regression analysis with two-stage cluster samples *Journal of the American Statistical Association* **88** 1388-1391.

Rao. J. N. K. and Thomas D. R. (2003) Analysis of categorical response data from complex surveys: an appraisal and update. In: Chambers R. and Skinner C. (eds) *Analysis of Survey Data* Chichester: Wiley.

Roberts G. ,Rao J. N. K. and Kumar S. (1987) Logistic regression analysis of sample survey data *Biometrika* **74** 1-12.

Rubin D. B(1987) *Multiple Imputation for Nonresponse in Surveys* New York: Wiley.

Rubin D. B. (1996) Multiple Imputation After 18 + Years *Journal of the American Statistical Association* **91** 473-489.

Särndal C. -E. (1996) For a better understanding of imputation. In Laaksonen S. (ed.) (1996). *International perspectives on nonresponse*. Proceeding of the sixth international workshop on household survey nonresponse. Helsinki: Statistics Finland, Research reports 219.

Särndal C. -E. (2001) Design-based methodologies for domain estimation. In: Lehtonen R. and Djerf K. (eds) *Lecture Notes on Estimation for Population and Domains Small Areas* Helsinki: Statistics Finland Reviews 2001/55-49.

Särndal C. -E. Swensson B. and Wretman J. (1992) *Model Assisted Survey Sampling* New York: Springer.

Satterthwaite F. E. (1946) An approximate distribution of estimates of variance components *Biometrics* **2** 110-114.

Schafer J. L. (2000) *Analysis of Incomplete Multivariate Data* New York: Chapman & Hall.

Schaible, W. L. (ed.) (1996) *Indirect Estimators in U. S. Federal Programs* New York: Springer.

Scott A. J. (1986) Logistic regression with survey data. *Proceedings of the Section on Survey Research Methods* American Statistical Association, 25-30.

Scott A. J. , Rao J. N. K. and Thomas D. R. (1990) Weighted least-squares and quasilikelihood estimation for categorical data under singular models *Linear Algebra and its Applications* **127** 427-447.

Shao J. and Tu D. (1995) *The Jackknife and Bootstrap* New York: Springer.

Silva P. L. N. and Skinner C. J. (1997) Variable selection for regression estimation in finite populations. *Survey Methodology* **23**, 23-32.

Singh A. C. (1985) On Optimal Asymptotic Tests for Analysis of Categorical Data from Sample Surveys Working Paper No. SSMD 86-002, Social Methods Division, Statistcs Canada.

Singh M. P. , Gambino J. and Mantel H. J. (1994) Issues and strategies for small area data *Survey Methodology* **20** 3-22.

Singh A. C. , Stukel D. M. and Pfeffermann D. (1998) Bayesian versus frequentist measures of error in small area estimation *Journal of the Royal Statistical Society, B* **60** 377-396.

Sitter R. R. (1992) A resampling procedure for complex survey data *Jouranal of the American Statistical Assoctiation* **87** 755-765.

Sitter R. R. (1997) Variance estimation for the regression estimator in two -phase sampling *Journal of the American Statistical Association* **92** 780-787.

Skinner C. J. ,Holt D. and Smith T. M. F. (eds) (1989) *Analysis of Complex Surveys* Chichester: Wiley.

Snijders T. A. B. and Bosker R. J. (2002) *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modelling* London: Sage Publications.

Sudman S. (1976) *Applied Sampling* New York: Academic Press.

Tepping B. J. (1968) Variance estimation in complex surveys *Proceedings of the Social Statistics Section* American Statistical Association 11-18.

Thomas D. R. and Rao J. N. K. (1987) Small-sample comparisons of level and power for simple goodness-of -fit statistics under cluster sampling *Journal of the American Statistical Association* **82** 630-636.

Thomas D. R. , Singh A. C. and Roberts G. R. (1996) Tests of independence on two-way tables under cluster sampling: an evaluation *International Statistical Review* **64** 295-311.

Valliant R. , Dorfman A. H. and Royall R. M. (2000) *Finite Population Sampling and Inference* New Yok: Wiley.

Verma V. , Scott C. and O' Muircheartaigh C. (1980) Sample designs and sampling errors for the World Fertility Survey *Journal of the Royal Statistical Society A* **143** 431-473.

Wald A. (1943) Tests of statistical hypotheses concerning several parameters when the number of observations is large *Transactions of the American Mathematical Society* **54** 426-482.

Williams D. A. (1982) Extra-binomial variation in logistic linear models *Applied Statistics* **31** 144-148.

Wilson J. R. (1989) Chi-square tests for overdispersion with multiparameter estimates *Applied Statistics* **38** 441-453.

Wolter K. M. (1985) *Introduction to Variance Estimation* New York: Springer.

Woodruff R. S. (1971) A simple method for approximating the variance of a complicated estimate *Journal of the American Statistical Association* **66** 411-414.

You Y. and Rao J. N. K. (2000) Hierarchical Bayes estimation of small area means using multi-level models *Survey Methodology* **26** 173-181.

You Y. and Rao J. N. K. (2002) A pseudo-empirical best linear unbiased prediction approach to small-area estimation using survey weights *The Canadian Journal of Statistics* **30** 431-439.

Yung W. and Rao J. N. K. (2000) Jackknife variance estimation under imputation for esti-mators using poststratification information *Journal of the American Statistical Association* **95** 903-915.

Ziegler A. , Kastner C. and Blettner M. (1998) The generalized estimating equations: an annotated bibliography *Biometrical Journal* **40** 115-139.